

NLP4DH 2024

**The 4th International Conference on
Natural Language Processing for Digital Humanities**

Proceedings of the Conference

November 16, 2024

©2024 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-181-0

Preface

Textual sources are a cornerstone for research in the digital humanities, especially when larger datasets are involved. In such cases, the use of natural language processing (NLP) technologies becomes indispensable. However, a significant challenge arises as NLP is often tailored to standard written languages, which can differ substantially from the genres and text types of primary interest to today's digital humanists. This complexity increases when research is conducted on minority languages, historical texts, or dialectal materials.

NLP, with its strong foundation in computer science, tends to focus on optimizing numerical outcomes and addressing abstract tasks like machine translation, poem generation, or sentiment analysis. In contrast, digital humanities, deeply rooted in the humanities, focuses on concrete research questions aimed at understanding specific phenomena rather than merely solving a problem. Where NLP strives to validate its methods, digital humanists often take methodological validity as a given, treating these tools as a means to an end rather than the end goal itself. This fundamental difference in approach highlights the need for more collaborative spaces where scholars from both fields can come together to exchange knowledge and methodologies.

Digital humanists are increasingly aware of the limitations of current NLP tools, while the NLP community has already developed several viable solutions. However, both fields would greatly benefit from enhanced communication. For instance, techniques like model fine-tuning and retraining, widely used in NLP, could be harnessed to improve results for diverse linguistic varieties. Similarly, research in the digital humanities frequently produces open datasets, offering opportunities for comparative analysis across different NLP strategies. This workshop aimed to stimulate such dialogue, encouraging the sharing of examples that illustrate how NLP tools can best address the research questions central to the humanities.

This year NLP4DH received 81 submissions which was our highest number yet.

Organizing Committee

- Mika Hämäläinen, Metropolia University of Applied Sciences
- Emily Öhman, Waseda University
- So Miyagawa, The University of Tsukuba / National Institute for Japanese Language and Linguistics
- Khalid Alnajjar, F-Secure Oyj
- Yuri Bizzoni, Aarhus University

Program Committee

- Joshua Wilbur, University of Tartu
- Stefania Degaetano-Ortlieb, Saarland University
- Luke Gessler, University of Colorado Boulder
- Leo Leppänen, University of Helsinki
- Quan Duong, University of Helsinki
- Iana Atanassova, University of Franche-Comté
- Won Ik Cho, Samsung
- Tyler Shoemaker, Dartmouth College
- Jouni Tuominen, University of Helsinki
- Enrique Manjavacas, Arevalo University of Leiden
- Kenichi Iwatsuki, Mirai Translate
- Matej Martinc, Jožef Stefan Institute
- Maciej Janicki, University of Helsinki
- Shuo Zhang, Bose
- Aynat Rubinstein, The Hebrew University of Jerusalem
- Frederik Arnold, Humboldt University of Berlin
- Thibault Clerice, National Institute for Research in Digital Science and Technology
- Nicolas Gutehrlé, University Bourgogne Franche-Comté
- Lama Alqazlan, University of Warwick
- Lidia Pivovarova, University of Helsinki
- Balázs Indig, Eötvös Loránd University
- Pierre Magistry, Institut national des langues et civilisations orientales
- Yoshifumi Kawasaki, The University of Tokyo

- Anna Dmitrieva, University of Helsinki
- Antti Kanner, University of Helsinki
- Maria Antoniak, Allen Institute for AI
- Katerina Korre, University of Bologna
- Daniela Teodorescu, University of Alberta
- Dongqi Pu, Saarland University
- Nils Hjortnaes, Indiana University Bloomington
- Noémi Ligeti-Nagy, Hungarian Research Centre for Linguistics
- Allison Lahnala, University of Bonn
- Gabriel Simmons, University of California, Davis
- Vilja Hulden, University of Colorado Boulder
- Jaihyun Park, Nanyang Technological University
- Jonne Sälevä, Brandeis University
- Martin Ruskov, University of Milan
- Youngsook Song, Sionix AI
- Pascale Moreira, Aarhus University
- Maciej Kurzynski, Lingnan University
- Aatu Liimatta, University of Helsinki
- Sourav Das, Indian Institute of Information Technology Kalyani
- Sebastian Oliver Eck, University of Music Franz Liszt Weimar
- Elissa Nakajima, Wickham Waseda University
- Nicole Miu Takagi, Waseda University
- Ken Kawamura, Revelata Inc
- Bo Dang, San Francisco Bay University
- Jack Rueter, University of Helsinki

Table of Contents

<i>Text Length and the Function of Intentionality: A Case Study of Contrastive Subreddits</i> Emily Sofi Ohman and Aatu Liimatta	1
<i>Tracing the Genealogies of Ideas with Sentence Embeddings</i> Lucian Li	9
<i>Evaluating Computational Representations of Character: An Austen Character Similarity Benchmark</i> Funing Yang and Carolyn Jane Anderson.....	17
<i>Investigating Expert-in-the-Loop LLM Discourse Patterns for Ancient Intertextual Analysis</i> Ray Umphrey, Jesse Roberts and Lindsey Roberts.....	31
<i>Extracting Relations from Ecclesiastical Cultural Heritage Texts</i> Giulia Cruciani	41
<i>Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience</i> Lucija Krusic.....	51
<i>It is a Truth Individually Acknowledged: Cross-references On Demand</i> Piper Vasicek, Courtni Byun and Kevin Seppi	63
<i>Extracting position titles from unstructured historical job advertisements</i> Klara Venglarova, Raven Adam and Georg Vogeler	75
<i>Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs</i> Natalie Hervieux, Peiran Yao, Susan Brown and Denilson Barbosa	85
<i>NLP for Digital Humanities: Processing Chronological Text Corpora</i> Adam Pawłowski and Tomasz Walkowiak.....	105
<i>A Multi-task Framework with Enhanced Hierarchical Attention for Sentiment Analysis on Classical Chinese Poetry: Utilizing Information from Short Lines</i> Quanqi Du and Veronique Hoste	113
<i>Exploring Similarity Measures and Intertextuality in Vedic Sanskrit Literature</i> So Miyagawa, Yuki Kyogoku, Yuzuki Tsukagoshi and Kyoko Amano.....	123
<i>Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction</i> Laura Manrique-Gomez, Tony Montes, Arturo Rodriguez Herrera and Ruben Manrique	132
<i>Canonical Status and Literary Influence: A Comparative Study of Danish Novels from the Modern Breakthrough (1870–1900)</i> Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos, Kenneth Enevoldsen, Katrine Baunvig and Kristoffer Nielbo.....	140
<i>Deciphering psycho-social effects of Eating Disorder : Analysis of Reddit Posts using Large Language Model(LLM)s and Topic Modeling</i> Medini Chopra, Anindita Chatterjee, Lipika Dey and Partha Pratim Das.....	156
<i>Topic-Aware Causal Intervention for Counterfactual Detection</i> Thong Thanh Nguyen and Truc-My Nguyen	165

<i>UD for German Poetry</i>	
Stefanie Dipper and Ronja Laarmann-Quante	177
<i>Molyé: A Corpus-based Approach to Language Contact in Colonial France</i>	
Rasul Dent, Juliette Janes, Thibault Clerice, Pedro Ortiz Suarez and Benoît Sagot	189
<i>Vector Poetics: Parallel Couplet Detection in Classical Chinese Poetry</i>	
Maciej Kurzynski, Xiaotong Xu and Yu Feng	200
<i>Adapting Measures of Literality for Use with Historical Language Data</i>	
Adam Roussel	209
<i>Improving Latin Dependency Parsing by Combining Treebanks and Predictions</i>	
Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala and Jenna Kanerva	216
<i>From N-grams to Pre-trained Multilingual Models For Language Identification</i>	
Thapelo Andrew Sindane and Vukosi Marivate	229
<i>Visualising Changes in Semantic Neighbourhoods of English Noun Compounds over Time</i>	
Malak Rassem, Myrto Tsigkouli, Chris W. Jenkins, Filip Miletic and Sabine Schulte im Walde	240
<i>SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek</i>	
Konstantin Schulz and Florian Deichsler	247
<i>A Two-Model Approach for Humour Style Recognition</i>	
Mary Ogbuka Kenneth, Foad Khosmood and Abbas Edalat	259
<i>N-gram-Based Preprocessing for Sandhi Reversion in Vedic Sanskrit</i>	
Yuzuki Tsukagoshi and Ikki Ohmukai	275
<i>Enhancing Swedish Parliamentary Data: Annotation, Accessibility, and Application in Digital Humanities</i>	
Shafqat Mumtaz Virk, Claes Ohlsson, Nina Tahmasebi, Henrik Björck and Leif Runefelt	280
<i>Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams</i>	
Roberts Dargis, Guntis Bārzdiņš, Inguna Skadiņa and Baiba Saulite	289
<i>Computational Methods for the Analysis of Complementizer Variability in Language and Literature: The Case of Hebrew "she-" and "ki"</i>	
Avi Shmidman and Aynat Rubinstein	294
<i>From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations</i>	
Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo and Veronika Laippala	308
<i>Testing and Adapting the Representational Abilities of Large Language Models on Folktales in Low-Resource Languages</i>	
J. A. Meaney, Beatrice Alex and William Lamb	319
<i>Examining Language Modeling Assumptions Using an Annotated Literary Dialect Corpus</i>	
Craig Messner and Thomas Lippincott	325

<i>Evaluating Language Models in Location Referring Expression Extraction from Early Modern and Contemporary Japanese Texts</i>	
Ayuki Katayama, Yusuke Sakai, Shohei Higashiyama, Hiroki Ouchi, Ayano Takeuchi, Ryo Bando, Yuta Hashimoto, Toshinobu Ogiso and Taro Watanabe	331
<i>Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction</i>	
Woori Jang and Seohyon Jung	339
<i>Text vs. Transcription: A Study of Differences Between the Writing and Speeches of U.S. Presidents</i>	
Mina Rajaei Moghadam, Mosab Rezaei, Gülşat Aygen and Reva Freedman	352
<i>Mitigating Biases to Embrace Diversity: A Comprehensive Annotation Benchmark for Toxic Language</i>	
Xinmeng Hou	362
<i>Classification of Buddhist Verses: The Efficacy and Limitations of Transformer-Based Models</i>	
Nikita Neveditsin, Ambuja Salgaonkar, Pawan Lingras and Vijay Mago	377
<i>Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora</i>	
Amanda Myntti, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala and Erik Henriksen	386
<i>Sui Generis: Large Language Models for Authorship Attribution and Verification in Latin</i>	
Svetlana Gorovaia, Gleb Schmidt and Ivan P. Yamshchikov	398
<i>Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration</i>	
Ryo Igarashi and So Miyagawa	413
<i>Exploring Large Language Models for Qualitative Data Analysis</i>	
Tim Fischer and Chris Biemann	423
<i>Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of RNNs, Transformers and LLMs</i>	
Chahan Vidal-Gorène, Nadi Tomeh and Victoria Khurshudyan	438
<i>Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference for Cost-Effective Cultural Heritage Dataset Generation</i>	
William Thorne, Ambrose Robinson, Bohua Peng, Chenghua Lin and Diana Maynard	450
<i>Assessing Large Language Models in Translating Coptic and Ancient Greek Ostraca</i>	
Audric-Charles Wannaz and So Miyagawa	463
<i>The Social Lives of Literary Characters: Combining citizen science and language models to understand narrative social networks</i>	
Andrew Piper, Michael Xu and Derek Ruths	472
<i>Multi-word expressions in biomedical abstracts and their plain English adaptations</i>	
Sergei Bagdasarov and Elke Teich	483
<i>Assessing the Performance of ChatGPT-4, Fine-tuned BERT and Traditional ML Models on Moroccan Arabic Sentiment Analysis</i>	
Mohamed HANNANI, Abdelhadi Soudi and Kristof Van Laerhoven	489

<i>Analyzing Pokémon and Mario Streamers' Twitch Chat with LLM-based User Embeddings</i>	
Mika Hämäläinen, Jack Rueter and Khalid Alnajjar	499
<i>Corpus Development Based on Conflict Structures in the Security Field and LLM Bias Verification</i>	
Keito Inoshita	504
<i>Generating Interpretations of Policy Announcements</i>	
Andreas Marfurt, Ashley Thornton, David Sylvan and James Henderson	513
<i>Order Up! Micromanaging Inconsistencies in ChatGPT-4o Text Analyses</i>	
Erkki Mervaala and Ilona Kousa	521
<i>CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration</i>	
Anton Eklund, Mona Forsman and Frank Drewes	536
<i>Empowering Teachers with Usability-Oriented LLM-Based Tools for Digital Pedagogy</i>	
Melany Vanessa Macias, Lev Kharlashkin, Leo Einari Huovinen and Mika Hämäläinen	549

Conference Program

16.11.2024

9:10–10:30 **Oral session 1**

9:10–9:30 *Lightning talks*

9:30–9:50 *Text Length and the Function of Intentionality: A Case Study of Contrastive Sub-reddits*

Emily Sofi Ohman and Aatu Liimatta

9:50–10:10 *Tracing the Genealogies of Ideas with Sentence Embeddings*

Lucian Li

10:10–10:30 *Evaluating Computational Representations of Character: An Austen Character Similarity Benchmark*

Funing Yang and Carolyn Jane Anderson

10:30–11:00 **Coffee break**

11:00–12:40 **Oral session 2**

11:00–11:20 *Investigating Expert-in-the-Loop LLM Discourse Patterns for Ancient Intertextual Analysis*

Ray Umphrey, Jesse Roberts and Lindsey Roberts

11:20–11:40 *Extracting Relations from Ecclesiastical Cultural Heritage Texts*

Giulia Cruciani

11:40–12:00 *Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience*

Lucija Krusic

12:00–12:20 *It is a Truth Individually Acknowledged: Cross-references On Demand*

Piper Vasicek, Courtni Byun and Kevin Seppi

12:20–12:40 *Extracting position titles from unstructured historical job advertisements*

Klara Venglarova, Raven Adam and Georg Vogeler

16.11.2024 (continued)

12:40–13:10 Lunch

13:10–15:30 Oral session 3

13:10–13:30 *Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs*

Natalie Hervieux, Peiran Yao, Susan Brown and Denilson Barbosa

13:30–13:50 *NLP for Digital Humanities: Processing Chronological Text Corpora*

Adam Pawłowski and Tomasz Walkowiak

13:50–14:10 *A Multi-task Framework with Enhanced Hierarchical Attention for Sentiment Analysis on Classical Chinese Poetry: Utilizing Information from Short Lines*

Quanqi Du and Veronique Hoste

14:10–14:30 *Exploring Similarity Measures and Intertextuality in Vedic Sanskrit Literature*

So Miyagawa, Yuki Kyogoku, Yuzuki Tsukagoshi and Kyoko Amano

14:30–14:50 *Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction*

Laura Manrique-Gomez, Tony Montes, Arturo Rodriguez Herrera and Ruben Manrique

14:50–15:10 *Canonical Status and Literary Influence: A Comparative Study of Danish Novels from the Modern Breakthrough (1870–1900)*

Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos, Kenneth Enevoldsen, Katrine Baunvig and Kristoffer Nielbo

15:10–15:30 *Deciphering psycho-social effects of Eating Disorder : Analysis of Reddit Posts using Large Language Model(LLM)s and Topic Modeling*

Medini Chopra, Anindita Chatterjee, Lipika Dey and Partha Pratim Das

16.11.2024 (continued)

15:30–16:30 Posters and coffee

- 15:30–16:30 *Topic-Aware Causal Intervention for Counterfactual Detection*
Thong Thanh Nguyen and Truc-My Nguyen
- 15:30–16:30 *UD for German Poetry*
Stefanie Dipper and Ronja Laarmann-Quante
- 15:30–16:30 *Molyé: A Corpus-based Approach to Language Contact in Colonial France*
Rasul Dent, juliette janes, Thibault Clerice, Pedro Ortiz Suarez and Benoît Sagot
- 15:30–16:30 *Vector Poetics: Parallel Couplet Detection in Classical Chinese Poetry*
Maciej Kurzynski, Xiaotong Xu and Yu Feng
- 15:30–16:30 *Adapting Measures of Literality for Use with Historical Language Data*
Adam Roussel
- 15:30–16:30 *Improving Latin Dependency Parsing by Combining Treebanks and Predictions*
Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala and Jenna Kanerva
- 15:30–16:30 *From N-grams to Pre-trained Multilingual Models For Language Identification*
Thapelo Andrew Sindane and Vukosi Marivate
- 15:30–16:30 *Visualising Changes in Semantic Neighbourhoods of English Noun Compounds over Time*
Malak Rassem, Myrto Tsigkouli, Chris W. Jenkins, Filip Miletić and Sabine Schultes im Walde
- 15:30–16:30 *SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek*
Konstantin Schulz and Florian Deichsler
- 15:30–16:30 *A Two-Model Approach for Humour Style Recognition*
Mary Ogbuka Kenneth, Foaad Khosmood and Abbas Edalat
- 15:30–16:30 *N-gram-Based Preprocessing for Sandhi Reversion in Vedic Sanskrit*
Yuzuki Tsukagoshi and Ikki Ohmukai

16.11.2024 (continued)

- 15:30–16:30 *Enhancing Swedish Parliamentary Data: Annotation, Accessibility, and Application in Digital Humanities*
Shafqat Mumtaz Virk, Claes Ohlsson, Nina Tahmasebi, Henrik Björck and Leif Runefelt
- 15:30–16:30 *Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams*
Roberts Dargis, Guntis Bārzdiņš, Inguna Skadiņa and Baiba Saulite
- 15:30–16:30 *Computational Methods for the Analysis of Complementizer Variability in Language and Literature: The Case of Hebrew "she-" and "ki"*
Avi Shmidman and Aynat Rubinstein
- 15:30–16:30 *From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations*
Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson, Anni Eskelinen, Liina Repo and Veronika Laippala
- 15:30–16:30 *Testing and Adapting the Representational Abilities of Large Language Models on Folktales in Low-Resource Languages*
J. A. Meaney, Beatrice Alex and William Lamb
- 15:30–16:30 *Examining Language Modeling Assumptions Using an Annotated Literary Dialect Corpus*
Craig Messner and Thomas Lippincott
- 15:30–16:30 *Evaluating Language Models in Location Referring Expression Extraction from Early Modern and Contemporary Japanese Texts*
Ayuki Katayama, Yusuke Sakai, Shohei Higashiyama, Hiroki Ouchi, Ayano Takeuchi, Ryo Bando, Yuta Hashimoto, Toshinobu Ogiso and Taro Watanabe
- 15:30–16:30 *Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction*
Woori Jang and Seohyon Jung
- 15:30–16:30 *Text vs. Transcription: A Study of Differences Between the Writing and Speeches of U.S. Presidents*
Mina Rajaei Moghadam, Mosab Rezaei, Gülşat Aygen and Reva Freedman
- 15:30–16:30 *Mitigating Biases to Embrace Diversity: A Comprehensive Annotation Benchmark for Toxic Language*
Xinmeng Hou

16.11.2024 (continued)

16:30–17:30 Virtual posters

- 16:30–17:30 *Classification of Buddhist Verses: The Efficacy and Limitations of Transformer-Based Models*
Nikita Neveditsin, Ambuja Salgaonkar, Pawan Lingras and Vijay Mago
- 16:30–17:30 *Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora*
Amanda Myntti, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala and Erik Henriksson
- 16:30–17:30 *Sui Generis: Large Language Models for Authorship Attribution and Verification in Latin*
Svetlana Gorovaia, Gleb Schmidt and Ivan P. Yamshchikov
- 16:30–17:30 *Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration*
Ryo Igarashi and So Miyagawa
- 16:30–17:30 *Exploring Large Language Models for Qualitative Data Analysis*
Tim Fischer and Chris Biemann
- 16:30–17:30 *Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of RNNs, Transformers and LLMs*
Chahan Vidal-Gorène, Nadi Tomeh and Victoria Khurshudyan
- 16:30–17:30 *Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference for Cost-Effective Cultural Heritage Dataset Generation*
William Thorne, Ambrose Robinson, Bohua Peng, Chenghua Lin and Diana Maynard
- 16:30–17:30 *Assessing Large Language Models in Translating Coptic and Ancient Greek Ostraca*
Audric-Charles Wannaz and So Miyagawa
- 16:30–17:30 *The Social Lives of Literary Characters: Combining citizen science and language models to understand narrative social networks*
Andrew Piper, Michael Xu and Derek Ruths
- 16:30–17:30 *Multi-word expressions in biomedical abstracts and their plain English adaptations*
Sergei Bagdasarov and Elke Teich
- 16:30–17:30 *Assessing the Performance of ChatGPT-4, Fine-tuned BERT and Traditional ML Models on Moroccan Arabic Sentiment Analysis*
Mohamed HANNANI, Abdelhadi Souidi and Kristof Van Laerhoven

16.11.2024 (continued)

- 16:30–17:30 *Analyzing Pokémon and Mario Streamers' Twitch Chat with LLM-based User Embeddings*
Mika Hämäläinen, Jack Rueter and Khalid Alnajjar
- 16:30–17:30 *Corpus Development Based on Conflict Structures in the Security Field and LLM Bias Verification*
Keito Inoshita
- 16:30–17:30 *Generating Interpretations of Policy Announcements*
Andreas Marfurt, Ashley Thornton, David Sylvan and James Henderson
- 16:30–17:30 *Order Up! Micromanaging Inconsistencies in ChatGPT-4o Text Analyses*
Erkki Mervaala and Ilona Kousa
- 16:30–17:30 *CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration*
Anton Eklund, Mona Forsman and Frank Drewes
- 16:30–17:30 *Empowering Teachers with Usability-Oriented LLM-Based Tools for Digital Pedagogy*
Melany Vanessa Macias, Lev Kharlashkin, Leo Einari Huovinen and Mika Hämäläinen

Text Length and the Function of Intentionality: A Case Study of Contrastive Subreddits

Emily Ohman
Waseda University
ohman@waseda.jp

Aatu Liimatta
University of Helsinki
aatu.liimatta@helsinki.fi

Abstract

Text length is of central concern in natural language processing (NLP) tasks, yet it is very much under-researched. In this paper, we use social media data, specifically Reddit, to explore the function of text length and intentionality by contrasting subreddits of the same topic where one is considered more serious/professional/academic and the other more relaxed/beginner/layperson. We hypothesize that word choices are more deliberate and intentional in the more in-depth and professional subreddits with texts subsequently becoming longer as a function of this intentionality. We argue that this has deep implications for many applied NLP tasks such as emotion and sentiment analysis, fake news and disinformation detection, and other modeling tasks focused on social media and similar platforms where users interact with each other via the medium of text.

1 Introduction

The relationship between the length, intentionality, register, genre, and emotion-associated word distributions in texts is a complex one. The genre often dictates the length for the sake of convention, such as with academic writing. The register (i.e. language as it is used in a specific situation and for a specific purpose) also affects text length, most obviously when there are platform limitations that impose maximum character counts such as SMS messages or tweets, or based on the general pace of the platform. Both of these effects are linked to and can artificially influence the intentionality and framing of a text. We define intentionality as a careful, self-curated production of text with a deliberate purpose or goal behind the communication.

Different text genres also force text producers to convey their message more succinctly and thus perhaps with more intentionality. Examples of this include works of literature, poetry, and political speeches where the evocation of specific

emotions in the reader is a desired effect achieved by carefully choosing the “right” words (see e.g. [Lipsitz, 2018](#); [Koljonen et al., 2022](#)). We can see similar evocation tactics on social media too with, for example, “rage bait”¹ posts. Despite the near-ubiquitous presence of rage bait on social media, only a handful of academic papers have explored the topic, and these studies have been from a disinformation perspective rather than from an NLP or linguistic perspective (see e.g. [Jennings-Roche, 2023](#); [Clem, 2023](#); [Jagayat and Choma, 2023](#); [Curato, 2021](#); [Johnston, 2024](#); [La Rocca, 2022](#)).

In this study, we explore the functions of length, intentionality, affect, and register in contrastive pairs of corpora on the same topic. As our data, we use subreddits (topic-specific discussion forums on Reddit) where at least two separate subreddits exist for the same topic and one is considered to be more serious or in-depth, and the other more general in nature.

We hypothesize that (1) the average length of a text (post or comment) is longer in the more serious, in-depth subreddits, (2) the language in more serious subreddits has more variability and lexical density, and (3) that positive words carry less information than negative words and therefore the more serious subreddits have less positive words than their general subreddit counterpart showing a different aspect of negativity bias ([Kanouse and Hanson Jr, 1987](#)).

This study also contributes to the discussion on best practices in how to work with the computational aspects of texts of varying lengths.

2 Background and Related Work

Many studies make off-handed mentions of the different nature of tweets as compared to other social media texts that are not artificially constrained by

¹*Rage bait* is when social media content is perceived as having been carefully constructed to induce a maximal negative emotional response in readers for the sake of engagement.

length to the same degree (Öhman, 2021a). Some mention the “informal language and expressive content such as emojis and hashtags” (Demszky et al., 2020), and others discuss the limited length, self-contained nature, and the helpfulness of emojis and hashtags for annotators of tweets in contrast with Reddit comments that tend not to include emojis and are often highly context and conversation dependent (Öhman et al., 2020).

For the most part, when social media message length is studied, the focus is on optimization for marketing purposes (Stephen et al., 2015) or crisis communication strategies (Ma and Yates, 2014). One interesting prior study looked into “perceived partner responsiveness” and found that longer messages were perceived as more intimate whether length was measured by counting tokens, characters, words or non-word characters (Freeman and Brinkley, 2014). In this study we focus on word count because we examine word choice beyond length as a metric.

Perhaps the most famous law in quantitative linguistics, Zipf’s law (Zipf, 1935, 1949) describes the distribution of sorted measures. In terms of corpus linguistics, Zipf’s law states that when the words in a corpus are ordered by frequency, the value of the n th entry is inversely proportional to n . This frequency can also be used to measure the length of words and more recent studies have extended Zipf’s law to state that information content causes word length to increase (Piantadosi et al., 2011).

Garcia et al. (2012) studied the distribution of positive and negative words and their frequency as they relate to information density and word frequency overall. Using small emotion lexicons (1034 entries for English) they found that word valence and frequency of use are related, and in particular that positive emotional content is more common than negative content. However, because positive words are more common, they point out that the relative rarity of negative words causes them to carry more information.

More recently Singh et al. (2023) used readability and other linguistic complexity metrics to show that the negativity bias holds true for Reddit data too meaning that negative emotions are associated with more complex texts than positive emotions. They continue on to show that current state-of-the-art transformers such as BERT have more trouble with the more complex texts and discuss the implications of this for the evaluation of emotion and

sentiment analysis models.

Intentionality is a very understudied concept in NLP. Here we use the term to mean a careful selection of words to achieve a desired affective reaction in the reader, i.e., affective rhetorical devices. Intentionality has mostly been researched as part of the field of rhetoric (see e.g. Bitzer, 1968; Burke, 1969) and in social psychology for example, to examine how other people determine whether an action or message was intentional, but some have looked at the rhetoric and affect of political speech: Teneva (2021) showed that emotional appeals in political discourse aimed at social solidarity, group identification, and shaping public opinion in the Internet news discourse. Their conclusion was that political emotions play an important role in modern argumentation.

In quantitative linguistics, varieties of language that are defined by their situation of use and communicative function are called registers. While there are many approaches to register analysis (e.g. Biber, 1988; Halliday and Matthiessen, 2013), the analysis of registers is generally founded on the idea that certain linguistic features are more or less well-suited to certain situational and functional concerns, and as such, registers tend to prefer and disprefer different sets of linguistic features depending on the situational and functional circumstances of the register. For instance, narrative registers might prefer features such as past-tense verb forms and third-person pronouns.

While text length is commonly recognized as a confounding factor for linguistic analyses, text length itself is rarely the object of study in linguistics. However, Liimatta (2022, 2023) analyzes comment length on Reddit from the point of view of register. These studies show that the length of a text is not determined at random even in contexts where the author can in principle write a text of any length on Reddit and other similar social media platform, when they are not constrained by genre conventions, publisher requirements or limitations of technology. Instead, Liimatta (2022, 2023) demonstrates that text length is closely linked with the idea of register. Just like different communicative functions and situational concerns prefer linguistic features that are well-suited for the situation, so too do different communicative functions prefer text lengths that are similarly well-suited for the situation. Furthermore, Liimatta (2022) goes on to show that the associations between text length and register are not all universal: within Reddit data,

Subreddit name	category	Speakers (users)	Utterances (comments)	Conversations (posts)	Comment:Post Ratio
startrek	general	111,119	2,215,516	110,183	20.11
DaystromInstitute	in-depth	18,811	505,171	17,308	29.19
Aviation	general	93,784	1,171,366	128,447	9.12
Flying	in-depth	38,822	1,261,216	65,024	19.40
AskHistory	general	17,646	78,932	14,441	5.47
AskHistorians	in-depth	193,943	2,065,764	327,340	6.31
wow	general	520,414	14,618,201	809,585	18.10
CompetitiveWoW	in-depth	14,487	126,323	8,861	14.26
wown00bs	beginner	12,452	101,138	14,456	7.00

Table 1: Overview of subreddits

many subreddits can differ in terms of the role that text length plays in them.

Of particular relevance to the present study are the findings by Liimatta (2022, 2023) that longer comments on Reddit have higher frequencies of linguistic features associated with higher information density, such as nominalizations, as well as ones associated with a more complex argument structure, such as infinitives and certain modal verbs; whereas the frequencies of features associated with non-edited, casual, “on-line” production tend to be higher in shorter comments, including features such as contractions, subordinator *that* deletion, first-person singular pronouns, and private verbs. These results suggest that, in general, longer comments tend to be more informationally dense and more carefully edited, as opposed to shorter comments, which tend to be more casual, less informationally dense, and less edited.

3 Data

We used the convokit Reddit corpus (Chang et al., 2020) to collect data from related but contrastive subreddits. We chose pairs of subreddits that are on the same topic but split into a general and a specific, typically more serious or professional, subreddit. The selection of the subreddits was made based on the authors’ knowledge of the topics and the subreddits, with both authors agreeing on each pair of subreddits. The criteria here vary slightly per subreddit; for example, we chose *r/startrek* as a general subreddit representing Star Trek discussions and *r/DaystromInstitute* which is a more serious subreddit dedicated to “Serious, in-depth discussion about *Star Trek.*”², the pairs *r/Aviation* and *r/flying* where aviation is for enthusiasts and flying for pilots, and *r/AskHistory* and *r/AskHistorians*

²From the community description of *r/startrek*

where the latter has much more stringent requirements for both posts and comments than the former.

We attempted to pick subreddit pairs of similar relative standing within pairs, however, there are many differences between our chosen pairs. All of the subreddits can be considered to be niche or nerdy in some aspects, including the more general ones and therefore they might not be different enough from each other for some of the pairs. We expect the most significant differences to be shown in the pair *r/startrek* and *r/DaystromInstitute* because the first in the pair is a fairly popular topic of discussion in the mainstream consciousness.

As a TV show, Star Trek has been one of the most long-lasting and successful franchises on TV since the 1960s and it is also easy for more casual viewers to form opinions and ask questions on the subreddits (Weldes, 1999; Pearson and Davies, 2014).

Compared to some of the other pairs, *r/aviation* and *r/flying* for example, *r/aviation* is meant for enthusiasts and *r/flying* for pilots. However, aviation enthusiasts are a more niche group to begin with compared to TV, and therefore it is unlikely that a large percentage of posts would be from the perspective of a more general public. Similarly, although *r/AskHistorians* is one of the most strictly moderated subreddits on reddit with stringent guidelines about what top-level posts must contain and *r/AskHistory* was established as a more lax alternative with fewer citation criteria and the like, the questions and discussions remain rather similar. Finally, we have the World of Warcraft (WoW)-related subreddits, that could be said to be somewhat niche despite their immense popularity because casual players, let alone the general public, are unlikely to be actively discussing the topic. We expect *r/wow* and *r/wownoobs* to be somewhat sim-

ilar simply because r/wow is so large it is going to catch a large majority of posts that would be more suited for r/wownoobs, but we expect r/wow and r/competitivewow to show more differences on the general to in-depth axis similar to the Star Trek and aviation-related subreddits.

The sizes of the corpora are not uniform and therefore the results were normalized by token count where applicable. Table 1 shows the number of tokens, unique usernames (speakers), conversations (posts), utterances (comments), and the ratio of comments to post to indicate how much engagement is typical for a post on each subreddit.

4 Method

For type-token ratio calculations, we used the same token count as for normalization and lemmatized type counts. We used SpaCy for tokenization and lemmatization. No other preprocessing steps were taken. To calculate the distribution of emotion-associated words, we use the NRC emotion intensity lexicon (Mohammad, 2018) which divides words into Plutchik’s 8 core emotion categories of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust and associates a value between 0 and 1 for the intensity of the emotion. Our method can therefore detect the difference between, e.g., annoyance, anger, and rage and be judged more akin to valence (Öhman, 2021b). Although we measure the prevalence of emotion-associated words in all eight Plutchik categories, we focus on Joy and Anger as the proxy emotions we expect to represent positive and negative in this context.

5 Results

The average post/comment length in the less general subreddits are consistently longer than in the more general ones (see table 2). For most pairs, the difference in length is double for the more serious subreddits, but in the case of the History subreddits the average length of posts in the more serious subreddit is almost 15 times longer. For the World of Warcraft subreddits, the general subreddit’s average post length is about half of the more serious one as expected, but about the same as the beginner one. Additionally, the same pattern holds with word length as well; while the difference in average word length is not large, it is consistently longer in the more serious subreddits by roughly .2 characters. For example, for *startrek* the median word length was 3 and for the *DaystromInstitute* 4.

Note that the average post length refers to all comments on posts, not just the original post. The difference in length is much larger if going by the original post only. We chose to merge the comments and posts here to show that the entire conversation on the in-depth subreddits is more complex, rather than just the starting post.

Because a higher type-token ratio (TTR) has been associated with registers with a higher information density (e.g. Biber, 1988), and as such we might expect the less general subreddits to have a higher TTR, we calculated the moving average type-token ratios (MATTR) for the subreddits. We chose MATTR over TTR since the lengths of the posts between the subreddit pairs were just so drastically different that for example, for r/startrek a typical post would only have about 70 words of which between 55 and 60 were unique making the TTR values very high (.75 to .8), whereas r/DaystromInstitute had lower TTR values typically between .35 and .45 because a typical post was between 800 and 1400 words of which 400-500 were unique. However, the MATTR values for all the subreddits in question was between .80 and .81 suggesting that MATTR was no better than TTR as a measure for information density or intentionality in our texts.

In table 2 we present the log-likelihood significances between *joy* and *anger*. The full log-likelihood (see table 3) and emotion word distributions (see table 4) are presented in the appendix.

6 Analysis and Concluding Discussion

We found that the average length of a text in posts and comments were longer in the more serious, in-depth subreddits at almost twice the length on average, with some significant deviations for the two History subreddits where the posts on the more serious subreddit were almost 15 times longer than on the less serious one. This finding supports the earlier results on the relationship between text length and register, which associated longer Reddit comments with linguistic features related to information density and more carefully edited content. The World of Warcraft subreddits differed here slightly with the post lengths on the general subreddit still about half the length of the more serious one (33.4 vs. 59.7), but with little difference between the competitive and beginner subreddits (57.4 vs. 59.7). This might be because the general subreddit is less specific and less moderated in terms of allowed

Subreddit name	category	Tokens	Avg. comment length	Avg. word length	Emotion (log-likelihood)
startrek	general	77,286,488	34.88	3.88	Joy+
DaystromInstitute	in-depth	38,072,122	75.37	4.08	Anger+
Aviation	general	27,775,314	23.71	3.99	Joy+, Anger+
Flying	in-depth	49,372,833	39.15	3.78	
AskHistory	general	4,319,093	23.90	4.18	Joy+, Anger+
AskHistorians	in-depth	150,197,577	351.89	4.26	
wow	general	515,071,857	33.39	4.01	Joy++
CompetitiveWoW	in-depth	7,541,880	59.70	4.30	Anger++
wown00bs	beginner	5,781,910	57.17	3.97	Joy-, Anger-

Table 2: Results

content, or that by asking a question on a beginner subreddit posters feel less self-conscious about asking what might be considered “stupid questions” on the main subreddit. Perhaps the niche nature of the discussion is not so much about expertise but about specificity. Interestingly the average post length on the main WoW subreddit and the main Star Trek subreddit was about the same (33.4 vs. 34.9) and on the general r/aviation and r/AskHistorians (23.7 vs. 23.9).

Our finding that the serious subreddits have both longer average comment lengths and features associated with higher information content, such as higher average word length, is also in line with the findings by Liimatta (2022, 2023) correlating longer comment length on Reddit with higher frequencies of linguistic features associated with densification of information, such as nominalizations, and shorter comments with features of more casual and personal, less carefully considered language, as well as Piantadosi et al. (2011) that information content causes word length to increase.

Another difference became apparent when sorting the posts on the subreddits by all-time top posts; the top posts in the general subreddits tended to contain images or videos rather than the long essay-like posts in the more serious subreddits, further highlighting the differences between the general and niche.

For the emotions, we did find that the posts in the more general subreddits tended to contain more joy-related words at higher intensities than their more in-depth counterparts. However, the case for anger-related words was not as clear-cut. The results indicate that r/DaystromInstitute posts tend to contain more anger than r/startrek and the same can be said for r/wow when compared to r/CompetitiveWoW, but for the others, it seemed that the general sub-

reddit posts contained more emotion-related words at higher intensities in general.

In conclusion, we have shown that the length of a post differs based on what we consider intentionality, that is, the in-depth nature of the text with more niche conversation significantly longer than more general conversation on the same topic. We were unable to show a difference in lexical variation due to the short messages in the general subreddits, but the results support the idea that general discussions are more positive than the more complex, longer texts in the in-depth subreddits.

7 Future Work

The complex relationships between the length, intentionality, register, genre, and emotion-associated word distributions in texts remains an interesting and useful area of study. For instance, while we have demonstrated an indirect relationship between emotion-associated word distributions and certain types of register variation, inasmuch as both of them are linked with text length, it would be fruitful to perform a deeper analysis of the connections between different register dimensions, emotion-associated words, and text length, to better understand this multifaceted problem.

In future work we hope to include part-of-speech and syntactic structures as features related to register and intentionality. We also hope to study additional subreddits and better utilize both fine-tuned LLMs and qualitative analysis.

Limitations

The emotions lexicons used for the analysis are not specifically designed for social media which could influence how well the labeled emotions in the lexicon correspond to the emotions in our data.

Furthermore, in the selection process of the subreddits, several options were excluded due to computational limitations. Future studies should include a more varied selection of subreddits and more manual validation of the results.

Ethics Statement

All data used in this study were obtained from public forums using convokit. We recognize that user-generated content can be subject to ethical concerns regarding privacy and consent. However, given the public and anonymous nature of Reddit posts and the fact that we limited our analysis to text-level features (length, register variation, and intentionality), without making inferences about individual users or their personal identities, any negative impact on any specific users or communities is mitigated.

The study was conducted with respect for the autonomy of online community members, and we acknowledge that online discourse can contain sensitive content. To mitigate any risks of harm, we avoided analyzing subreddits that could contain vulnerable populations or sensitive topics.

Finally, the potential biases in our study, including selection bias or platform-specific biases, were considered in our methodology. We took care to transparently report these limitations in our findings, ensuring that our conclusions are contextualized within the broader ethical and social considerations of using online data.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K21058.

References

- Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.
- Lloyd F Bitzer. 1968. The rhetorical situation. *Philosophy & rhetoric*, pages 1–14.
- Kenneth Burke. 1969. A rhetoric of motives. *U of California P*.
- Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. *Convokit: A toolkit for the analysis of conversations*. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.
- Dany Clem. 2023. *Accessible and Activist Rhetorics: TikTok as a Learning Tool*. Ph.D. thesis, Arkansas State University.
- Nicole Curato. 2021. Interruptive protests in dysfunctional deliberative systems. *Politics*, 41(3):388–403.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. *Goemotions: A dataset of fine-grained emotions*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Linda Kramer Freeman and Jason Brinkley. 2014. *Length matters: Message metrics that result in higher levels of perceived partner responsiveness and changes in intimacy as friends communicate through social network sites*. *The Journal of Social Media in Society*, 3(1).
- David Garcia, Antonios Garas, and Frank Schweitzer. 2012. *Positive words carry less information than negative words*. *EPJ Data Science*, 1(1):1–12.
- M. A. K. Halliday and Christian M. I. M. Matthiessen. 2013. *Halliday’s Introduction to functional grammar*, 4 edition. Routledge, London.
- Arvin Jagayat and Becky L Choma. 2023. A primer on open-source, experimental social media simulation software: Opportunities for misinformation research and beyond. *Current Opinion in Psychology*, page 101726.
- Allison Jennings-Roche. 2023. Delegitimizing censorship: Contending with the rhetoric of an anti-democratic movement. *The Political Librarian*, 6(1).
- Hank Johnston. 2024. The maga movement’s big umbrella. *Mobilization: An International Quarterly*, 28(4):409–433.
- David E Kanouse and L Reid Hanson Jr. 1987. Negativity in evaluations. In *Attribution: Perceiving the cause of behavior*. Lawrence Erlbaum Associates, Inc.
- Juha Koljonen, Emily Öhman, Pertti Ahonen, and Mikko Mattila. 2022. *Strategic sentiments and emotions in post-second world war party manifestos in finland*. *Journal of computational social science*, 5(2):1529–1554.
- Gevisa La Rocca. 2022. The mediatization of disinformation as a social problem: The role of platforms and digital media ecology. *Information Disorder: Learning to Recognize Fake News*, Peter Lang, Berlin, pages 43–62.
- Aatu Liimatta. 2022. *Do registers have different functions for text length? A case study of Reddit*. *Register Studies*, 4(2):263–287.
- Aatu Liimatta. 2023. *Register variation across text lengths: Evidence from social media*. *International Journal of Corpus Linguistics*, 28(2):202–231. Publisher: John Benjamins.

- Keena Lipsitz. 2018. Playing with emotions: The effect of moral appeals in elite rhetoric. *Political Behavior*, 40(1):57–78.
- Xin Ma and Justin Yates. 2014. [Optimizing social media message dissemination problem for emergency communication](#). *Computers & Industrial Engineering*, 78:107–126.
- Saif Mohammad. 2018. [Word affect intensities](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Emily Öhman. 2021a. *The language of emotions: Building and applying computational methods for emotion detection for English and beyond*. UniGrafiya: University of Helsinki.
- Emily Öhman. 2021b. [The validity of lexicon-based sentiment analysis in interdisciplinary research](#). In *Proceedings of the workshop on natural language processing for digital humanities*, pages 7–12.
- Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. [Xed: A multilingual dataset for sentiment analysis and emotion detection](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552.
- Roberta Pearson and Máire Messenger Davies. 2014. *Star Trek and American Television*. Univ of California Press.
- Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.
- Pranaydeep Singh, Luna De Bruyne, Orphee De Clercq, and Els Lefever. 2023. [Misery loves complexity: Exploring linguistic complexity in the context of emotion detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12871–12880.
- Andrew T Stephen, Michael Sciandra, and Jeffrey Inman. 2015. Is it what you say or how you say it? how content characteristics affect consumer engagement with brands on facebook. *How Content Characteristics Affect Consumer Engagement with Brands on Facebook (October 1, 2015)*. Saïd Business School WP, 19.
- Ekaterina V Teneva. 2021. The rhetoric of political emotions in the internet news discourse. *Galactica Media: Journal of Media Studies*, 3(1):125–145.
- Jutta Weldes. 1999. Going cultural: Star trek, state action, and popular culture. *Millennium*, 28(1):117–134.
- George Kingsley Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.
- George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

Appendices

Log-likelihood

	subreddit	O1	1%	O2	2%	LL	%DIFF	Bayes	ELL	RRisk	LogRatio	OddsRatio
JOY	st/DI	1171246	1.52	421554	1.11	31983.22	36.87	31964.65	0.00002	1.37	0.45	1.37
ANG		619212	0.8	316266	0.83	272.86	-3.55	254.3	0	0.96	-0.05	0.96
JOY	av/fly	297945	1.07	559895	1.13	604.19	-5.41	586.02	0	0.95	-0.08	0.95
ANG		162804	0.59	225974	0.46	5700.07	28.07	5681.91	0.00001	1.28	0.36	1.28
JOY	askhist/hists	44201	1.02	1466122	0.98	94.65	4.84	75.8	0	1.05	0.07	1.05
ANG		42566	0.99	1221812	0.81	1427.85	21.15	1409	0	1.21	0.28	1.21
ANG		83855	0.51	594570	0.54	303.83	-6.18	285.17	0	0.94	-0.09	0.94
JOY	wow/comp	7329450	1.42	87558	1.16	3834.93	22.57	3814.86	0	1.23	0.29	1.23
ANG		5859108	1.14	112940	1.5	7683.63	-24.04	7663.55	0	0.76	-0.4	0.76
JOY	wow/noob	7329450	1.42	79103	1.37	123.34	4.01	103.27	0	1.04	0.06	1.04
ANG		5859108	1.14	49113	0.85	4581.69	33.92	4561.62	0	1.34	0.42	1.34

Table 3: Log-likelihood calculations

Emotion word distributions

	Nanger	Nanticipation	Ndisgust	Nfear	Njoy	Nsadness	Nsurprise	NTrust
startrek	80.11904	137.4671	46.7845575	97.41404086	151.546	77.83572	40.96666	236.8187
DaystromInstitute	83.07029	114.2822	36.260473	117.9215713	110.7251	75.66423	36.94725	220.891
aviation	58.61455	104.6397	32.8963086	97.75651717	107.2698	65.30503	45.19683	187.3386
flying	45.76897	126.2415	28.0944731	79.7299766	113.4014	58.57647	39.32689	206.2676
AskHistory	98.554	100.7872	48.1698912	157.3611404	102.3396	89.99936	38.91132	204.3497
AskHistorians	81.34697	95.62274	37.160274	129.5403759	97.61289	74.19905	34.24224	205.5677
wow	113.7532	126.5027	53.5203207	139.9568105	142.2995	94.2153	62.13591	211.5859
CompetitiveWoW	149.7507	120.5169	53.4918681	166.0155359	116.0964	111.2234	71.68543	198.8933
wownoob	84.94327	137.9898	33.5502715	116.2166153	136.812	72.26706	54.0252	226.2408

Table 4: Normalized Emotion-word distributions by intensity measures

Tracing the Genealogies of Ideas with Sentence Embeddings

Lucian Li

PhD Student

School of Information Science

University of Illinois, Urbana-Champaign

zilul2@illinois.edu

Abstract

Detecting intellectual influence in unstructured text is an important problem for a wide range of fields, including intellectual history, social science, and bibliometrics. A wide range of previous studies in computational social science and digital humanities have attempted to resolve this through a range of dictionary, embedding, and language model based methods.

I introduce an approach which leverages a sentence embedding index to efficiently search for similar ideas in a large historical corpus. This method remains robust in conditions of high OCR error found in real mass digitized historical corpora that disrupt previous published methods, while also capturing paraphrase and indirect influence.

I evaluate this method on a large corpus of 250,000 nonfiction texts from the 19th century, and find that discovered influence is in line with history of science literature. By expanding the scope of our search for influence and the origins of ideas beyond traditional structured corpora and canonical works and figures, we can get a more nuanced perspective on influence and idea dissemination that can encompass epistemically marginalized groups.

1 Introduction

In Darwin's Plots, (Beer, 2009) examines Darwin's influence on literature as a complex and reciprocal system. Beer identifies in Darwin's writings not only the influence of naturalists and geologists like Lyell, but also the stylistic and lyrical influence of Wordsworth, Coleridge, and Milton. Proceeding onwards, Beer delves into a close reading of how Darwinian metaphors, themes, and worldviews emerge in the works of George Eliot and Thomas Hardy, both correspondents of Darwin who wrote extensive commentaries and reactions to the Origin of Species.

As Beer's work shows, there are connections between intellectual figures and avenues for the

spread of ideas not possible to observe except through deliberately interdisciplinary efforts. But scholars cannot have expertise in every field and every potential author; experts with training in dozens of subfields and time to read hundreds of thousands of books are in short supply.

Computational methods can enable analysis across some of these boundaries. In this paper, I present a novel method to detect intellectual influence across a large corpus. Taking advantage of the unique affordances of large language models in encoding semantic and structural meaning while remaining robust to paraphrasing, we can search for substantively similar ideas and hints of intellectual influence in a computationally efficient manner. Such a method allows us to operationalize different levels of confidence: we can allow for direct quotation, paraphrase, or speculative similarity while remaining open about the limitations of each threshold.

I apply an ensemble method combining General Text Embeddings (GTE), a state-of-the-art sentence embedding method described in (Li et al., 2023) optimized to capture semantic content while also retaining aspects of style and vocabulary choice. I vectorize sentences from a corpus of roughly 250,000 nonfiction books and academic publications from the 19th century for instances of ideas and arguments appearing in Darwin's publications. This functions as an initial evaluation and proof of concept; the method is not limited to detecting Darwinian ideas but is detecting similarities on a large scale in a wide range of corpora and contexts

2 Related Literature

Previous attempts to quantify and detect intellectual influence have taken three overall directions: topic modelling, text reuse detection, and word sense similarity. Studies using topic models generally compare topic distributions across documents

or subdocuments. They can capture a zeitgeist of themes and shifting focus but lack granular focus on specific claims. (Rockmore et al., 2018) uses topic models to trace the genealogy of national constitutions. In (Barron et al., 2018), the authors measure K-L divergence of the Topic Distributions of French Revolutionary speeches. In general, these approaches are generally more effective in a limited context with a controlled set of topics and a high likelihood of influence between documents in the corpus. However, changes in topic distribution may reflect high level shifts in societal context rather than direct influence.

Text reuse methods focus on high confidence detection of exact quotation. They can detect one form of direct influence with near certainty but are more limited to paraphrasing and indirect influence. (Funk and Mullen, 2018) and (Smith et al., 2015) both search a large corpus for direct quotations while using a mix of computationally intensive corrections to remain robust to OCR errors. While direct quotation detection ensures high confidence, it necessarily only captures a very limited range of potential influence, excluding similarities in language use, indirect quotation, and similar claims. The n-gram alignment problem is also highly computationally intensive, and requires extensive resources to apply to large corpora.

Finally, approaches focused on detecting similarity and changes in word sense (for example, comparing diachronic embeddings of how concepts like ‘justice’ evolved over time) can capture stylistic and discursive influence. (Soni et al., 2021) studies Abolitionist newspapers uses word2vec word embeddings. Other approaches, such as (Vicinanza et al., 2023) use language models such as BERT to measure stability and innovation in word senses. However, these findings can be very difficult to interpret across entire vocabularies and are unable to capture any changes in content or argumentation. The influence they capture is also highly speculative; stylistic changes may reflect wider shifts in language use instead of direct interactions.

My proposed method attempts to synthesize text reuse and word sense embedding methods. By evaluating claims on the sentence level, we can gain a granular understanding of specific ideas, while also remaining open to abstract similarities in meaning or structure. Specialized sentence embeddings language models have demonstrated improved effectiveness in encoding semantic meaning in general evaluation tasks as compared to standard BERT and

Word2Vec embeddings (Reimers and Gurevych, 2019). Sentence embeddings have been applied to the task of detecting citation and plagiarism in general academic literature in (Alvi et al., 2021) and (Lagopoulos and Tsoumakas, 2021) as well as encoding documents specific to disciplinary subfields in (Chen et al., 2019). I selected GTE vectorization because of the lower computational demands of the GTE-small model and its higher performance in evaluation metrics to other sentence embedding methods.

Finally, the subword tokenization strategy used by BERT and more recent language models was demonstrated in (Nguyen et al., 2020) to be resilient to OCR error. Real large scale historical datasets, such as HathiTrust’s digitized book collection, have extensive OCR error, averaging 7% and up to 20% character error depending on the scan quality and time period (Jiang et al., 2021).

Previously published word embedding and topic model approaches are heavily impacted by OCR character error, and while some text reuse approaches mitigate OCR error through machine learning correction, these tend to be highly corpus specific.

3 Dataset

To evaluate my method, I constructed a dataset based around authors active in 19th century academic societies in the British Empire. I curated a list of journals based on secondary readings (Pal, 2014) (Barton, 1990) as well as prior knowledge about the period. This is not meant to capture comprehensively all academic publications in the 19th century, but rather to gather a representative cross section of the most active members of this community. Below is a list of the journals scraped:

- General:
 - Royal Society
 - Royal Institution
 - Cambridge Philosophical Society
- Chemical:
 - (London) Chemical Society
- Medical:
 - (London) Medical and Chirurgical Society
- Biological:

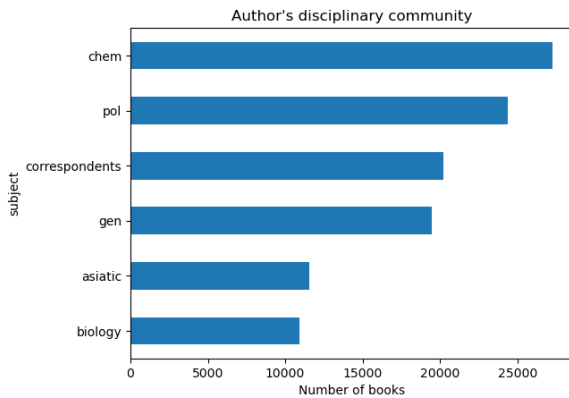


Figure 1: Distribution of disciplinary community of author (extensive overlaps between classes)

- Linnean Society
- Zoological Society
- Entomological Society
- Geographical:
 - Geographical Society
 - (Royal, Calcutta, American) Oriental Society
- Political and social scientific:
 - The Economist
 - Westminster Review
 - Edinburgh Review

I grouped these societies into proto-disciplines such as biology, geology, chemistry, and politics/social science. I constructed a supplementary dataset of books by Darwin's correspondents using letters from the Darwin Correspondence Project.¹ Author names were extracted from downloaded proceedings using Spacy's NER utility. 250,000 books by the 1,000,000 identified potential authors were downloaded as digitized texts from the Internet Archive and Project Gutenberg. Metadata about the books used are available in this csv.² I also used the Project Gutenberg editions of Darwin's Origin of Species and Descent of Man and Herbert Spencer's Principles of Sociology and Principles of Biology for a comparative sample.

4 Method

4.1 Preprocessing

I performed sentence tokenization per book using NLTK. Overly short documents (<1000 characters)

¹<https://www.darwinproject.ac.uk/>

²<https://uofi.app.box.com/file/1412863623947>

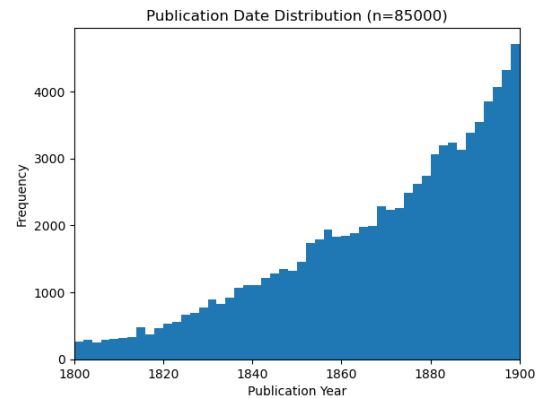


Figure 2: Distribution of books by year

and sentences (<45 characters) were removed because short documents tended to be either mislabelled, or consist mostly of images that could not be accurately converted to text. Short sentences tended to not contain enough information for a coherent argument, or represented formatting, index, and table of contents elements. No further text cleaning was performed; the BERT base of the model used for vectorization uses case and punctuation markings to encode meaning.

4.2 Model selection and finetuning

I used the General Text Embeddings (GTE) model, a BERT based approach fine-tuned with internet text and specific entailment tasks to capture semantic meaning. For a preliminary set of randomly sampled books, GTE embeddings were generated for each sentence using the GTE-small model implemented in the sentence-transformers Python package. GTE-small was selected due to memory and computational power constraints.

For fine tuning, I randomly sampled pairs of books to generate 1,000 pairs of sentence similarity scores. I inspected the pairs to label the accuracy of the score. If the sentences were similar due to purely coincidental factors (for example, transitional phrases like "I go on to argue" or "it should be obvious"), I assign a score of -1. If they have a missed similarity (i.e. making the same argument) but have a score that does not meet the threshold, they are assigned a score of 1. Otherwise, if the score is correct, the fine tuning score was left the same. It is difficult to determine the effectiveness of this method in resolving false positive matches across the broader corpus due to the lack of labelled data, but it successfully removed all instances of the hand annotated false positives from future matches.

I used these pairs to fine tune the GTE-small model using cosine embedding loss in Hugging Face.

4.3 Search

No additional hyperparameter changes were performed. Using my fine tuned version of the GTE-small model, I generated sentence embeddings of each sentence in the corpus. From these vectors, I used FAISS (Douze et al., 2024) to create rapidly searchable cosine indices for every sentence in the corpus. For further analysis, I used thresholds of >0.85 cosine similarity (speculative and low confidence), >0.90 cosine similarity (indirect/medium influence) and >0.95 (high confidence and direct quotation). All code for the project are available in this GitHub repository.³

5 Findings

5.1 Robustness to OCR error

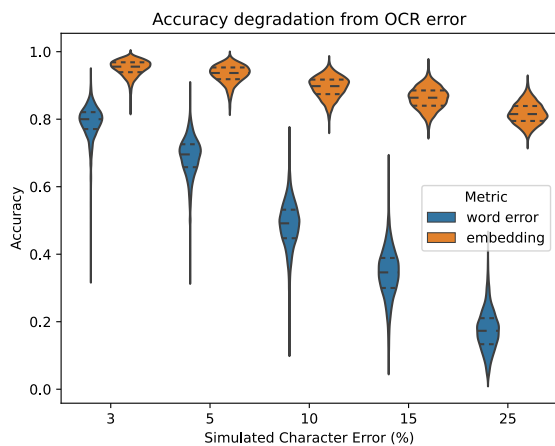


Figure 3: Error rates on simulated corrupted data with artificially permuted characters. The blue distributions show the word error rate at different levels of character error. The orange shows the cosine similarity in sentence embeddings between the original and corrupted text.

First, I evaluate the impact of low quality OCR on the performance of this method. Figure 3 demonstrates its robustness. I took a subset of the corpus consisting of human transcribed books from Project Gutenberg and simulated character error by randomly permuting parts of each sentence with a random character, integer, or empty string. I compared the sentence embedding representation of this new string with that of the original string, and

³<https://github.com/lucianli123/darwin-novelty>

found that there was generally very little decrease in cosine similarity. Even at 10% character error, roughly 90% of sentences will still be captured above the 0.85 cosine similarity threshold.

Conversely, word accuracy is highly sensitive to increased character error. At the most frequent CER of 7% in real scanned corpora, between 30-50% of words are corrupted. At higher CER levels, the overwhelming majority of words are lost. For dictionary based approaches, like text reuse, word embeddings, and topic modelling, this creates extensive accuracy issues.

This method is more robust to OCR error than dictionary based approaches at all CER levels, suggesting that in applications where transcription error is expected, this method will generally preserve more signal accuracy.

5.2 Validation against historical ground truth

Because annotated data does not exist for the very messy corpus of scanned 19th century books, I conducted evaluation against historical ground truth. I take the set of sentence embeddings for selected books by Darwin, Herbert Spencer, and 2 randomly sampled books published in the same year as *Origin of Species*. Based on academic consensus about Darwin (Mayr, 1995), and documented evidence about his correspondents, we would expect Darwin’s publications to display more similarity with his intellectual circle and in certain disciplinary communities (geology, natural history) vs (chemistry and orientalist circles).

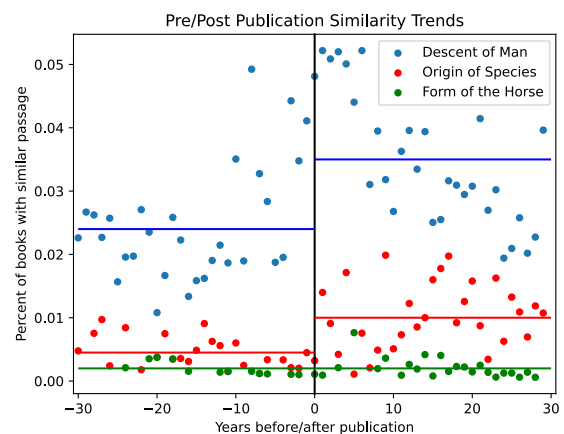


Figure 5: Similarity with books published before and after.

Also, we would expect the impact of Darwin’s books to display prescience, i.e., that they exert more influence on future publications than they re-

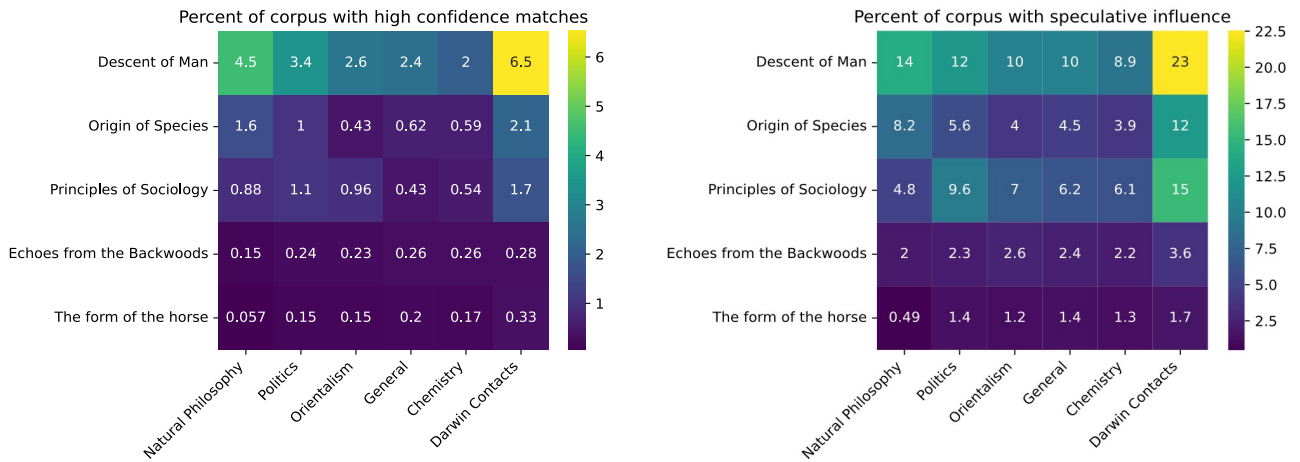


Figure 4: Percent of books post-publication with any detected influence. The last two books are randomly sampled books published in the same year as *Origin of Species* and included as a baseline comparison.

ceive from past ones. This would be consistent with the idea that Darwin’s publications revolutionized attitudes and ideas.

When we plot the influence over time (Figure 4), we see the method’s sensitivity to shifts in the overall discourse. Each point represents the similarity of books published each year and the colored lines represented the average similarity of all pre- and post- publication books. In red, the *Origin of Species* (1859) draws from a handful of primarily geological and biological sources prepublication, but radically shifts the overall discourse. In blue, the *Descent of Man* (1871) engages more with discourses across a diverse range of disciplines as well as the evolutionary ideas already introduced in the *Origin*. *Descent* has relatively more connections to previously published works, coming from Darwin’s main thesis already existing in the discourse community. However, it likewise radically shifts the discourse in the corpus. Both of Darwin’s major works proved innovative, as they drew less from previously published texts while exerting significant influence on future texts. The randomly sampled book in green shows that the effect is not likely due to corpus wide factors.

We can get a more detailed view of influence in specific disciplinary communities in Figure 5. The rightmost column of both Figure 5 heatmaps show the overrepresentation of Darwinian influence in books by Darwin’s correspondents (people with documented interactions with Darwin). As a further confirmation, we can see more influence from Darwin’s books in Biology and Geology than Chemistry or Political Theory. Even when the

confidence threshold is lowered and more speculative matches are allowed, the same patterns persist. While this may not give us confidence that all examples of influence are being detected, we can at least be more sure that the distributions of detected influence reflect some kind of underlying historical pattern.

The very low levels of influence detected for the "control" books in Figures 4 and 5 gives us some confidence in the resilience of this method against excessive false positives. Individual false positive matches do not result in a book level false signal, as books we expect to be obscure have extremely few matches across the corpus.

5.3 Close reading

In table 1, we see examples of sentences at each similarity threshold. In the first example, we can see that the method detects direct quotation at high confidence while remaining robust to OCR errors and minor structural and punctuation changes. The second example shows the ability of the method to identify cases of paraphrase with very limited shared word use. These two statements make the same claim, but only share a limited number of words. Corpus based approaches will likely fail to capture the similarity in ideas in this case.

Lastly, we see the ability of the method to capture speculative matches across genres. The first quote is from *Origin of Species* and the second book is from George Eliot’s *Middlemarch*. In this quote, we can see Darwin’s metaphor of the web of life echoed in *Middlemarch*. Eliot uses the same metaphor to describe a complex network of human

Sentence 1	Sentence 2	Cosine Similarity
Would it be believed, that the larvae of an insect, or fly, no larger than a grain of rice, destroy some thousand acres of pine-trees, many of them from two to three feet in diameter, and a hundred and fifty in height?	Would it be believed, says Wilson, the ornitholog-ist, ' that the larvs of an insect, or fly, no larger tliaan a grain of rice, should, destroy some thousand ncrees of pine trees, many of uiem two or three feet in diameter, and one himdred and fifty feet high.	High confidence (0.97)
I have called this principle, by which each slight variation, if useful, is preserved, by the term natural selection, in order to mark its relation to man's power of selection.	The expression "natural selection" was chosen as serving to indicate some parallelism with artificial selection—the selection exercised by breeders.	Medium confidence (0.92)
I have so much to do in unraveling certain human lots, and seeing how they were woven and interwoven, that all the light I can command must be concentrated on this particular web, and not dispersed over that tempting range of relevancies called the universe.	We shall never disentangle the inextricable web of affinities between the members of any one class; but when we have a distinct object in view, and do not look to some unknown plan of creation, we may hope to make sure but slow progress.	Speculative influence (0.85)

Table 1: Selected examples of sentence pairs with similarity scores

relationships. She draws the same conclusion as Darwin: that in a highly complex situation, we must focus on the particular rather than the general. We know this isn't random chance because literary scholars like Beer have examined the correspondence between Eliot and Darwin, but that relies on a whole infrastructure of experts in the papers of both authors. countless lesser known examples that subject area specialists haven't focused on studying remain unknown. The speculative matches in discovered here include extensive false positives, cases where stylistic or structural similarities don't suggest true influence, but may allow for the discovery of previously unknown influence.

6 Limitations

Because of the training process for GTE, semantic similarity is the main component in calculating embeddings. This captures the spread and influence of specific claims, but is much weaker in terms of metaphor, stylistic similarity, and influence in argument structure and construction.

For future work, I plan to create an ensemble approach generating AMR graphs (Opitz et al., 2021) or knowledge graphs from the structure of each individual argument. Then, graph embeddings can be generated through a neural network based approach like (Wang et al., 2018).

There are also issues with false positive matches,

particularly in terms of generic and stock sentences used as transitions or argumentative signposts rather than conveying a specific claim. First, once larger scale results are discovered across the corpus, highly frequent sentences across the corpus can simply be removed. I plan to train a relatively simple BERT based model to detect false positive matches, especially because there are commonly appearing stock phrases that account for a large percentage of false positives.

However, false negatives are likely impossible to adequately evaluate or completely remove. To identify with complete confidence all instances of false negatives, the entire corpus must be examined and annotated. The preliminary results presented in this paper suggest that the matches discovered by the method roughly approximate what we expect from historical research. While this is no guarantee against false positives and negatives, it suggests that the proportion of false negatives and positives is not dramatically skewed. But we must remain aware that this method is not able to comprehensively identify all influence, but instead discovers previously unknown avenues of research.

7 Conclusion

7.1 Future directions

This method allows for a hypertextual exploration of any given text. As shown in Figure 6, it can

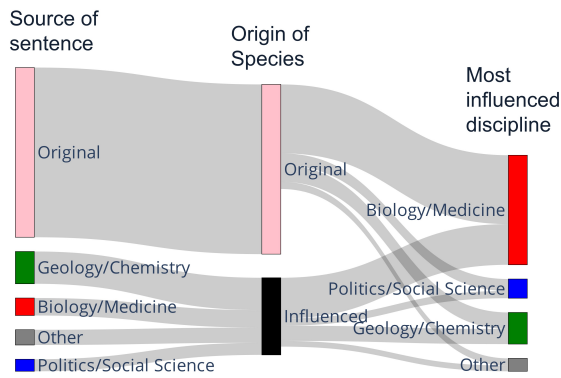


Figure 6: Origins and post-publication influence of each statement in *Origin of Species*

trace all occurrences of a specific claim over time in the corpus. Imagine an edition of the *Origin of Species* where a reader can click each sentence and receive information on where that argument appeared pre-publication. They would be able to observe the heavy influences from geology, as well as Darwin’s own original observations based on his travels. A reader would then be able to look forward and see which fields each statement resonated with and the context for how they read and interpreted sections differently – in our Darwinian case, similarities and differences – in ways eugenicists read the *Origin* compared to botanists. I hope to collaborate and make these enhanced editions available to a wide range of humanist scholars. By enabling researchers to perform more comprehensive searches for the origins and impacts of claims in their subjects and texts of interests, I hope to open additional avenues for interesting research.

Now imagine this on a larger scale: instead of arguments from the *Origin*, all arguments in the corpus. Would we be able to find common features of ideas which gained wider traction or leapt across disciplinary boundaries? My future work will focus on larger scale patterns in this corpus, with particular focus on the generalizable qualities of ideas or authors whose ideas gained influence beyond their disciplinary communities.

7.2 Reflection

Traditional narratives of discovery and invention valorize the contributions of individual geniuses – almost exclusively wealthy men from metropolitan societies. While historians of science have challenged this paradigm, the types of sources currently available for historical resource have limited practical moves toward reform. Dependence on personal

papers and close reading of related works limits the potential scale and representativeness of these efforts; at some point, it becomes impossible to read the hundreds of thousands of now unknown publications. Even Beer’s incisive work ultimately limits itself to Anglo-American literature and canonical authors. Responsible use of potentially destabilizing new AI technologies, keeping in mind their gaps and exclusions, can radically reshape our view of genealogies of ideas and influence and suggest previously unexplored possibilities for further exploration.

This mode of analysis has the potential to uncover connections between the work of hundreds of thousands of authors, among them women explorers and scientists, interlocutors from colonized peoples, and simply those whose ideas and contributions have been forgotten in the present. These ideas are as much part of the patchwork of intellectual life in the 19th century as those of Darwin or Herbert Spencer or Charles Lyell. Taking a wider view has the potential to reinvent the history of science.

8 Acknowledgements

Thank you to Professor Ted Underwood for discussion about the intellectual goals and evaluation metrics of this project, as well as help reading and editing earlier drafts.

Thank you as well to the anonymous reviewers for helpful comments and points of clarification.

References

- Faisal Alvi, Mark Stevenson, and Paul Clough. 2021. Paraphrase type identification for plagiarism detection using contexts and word embeddings. *International Journal of Educational Technology in Higher Education*, 18(1):42.
- Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.
- Ruth Barton. 1990. ‘an influential set of chaps’: The x-club and royal society politics 1864–85. *The British journal for the history of science*, 23(1):53–81.
- Gillian Beer. 2009. *Darwin’s plots: Evolutionary narrative in Darwin, George Eliot and nineteenth-century fiction*. Cambridge University Press.
- Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for

- biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. *The faiss library*. *Preprint*, arXiv:2401.08281.
- Kellen Funk and Lincoln A Mullen. 2018. The spine of american law: Digital text analysis and us legal practice. *The American Historical Review*, 123(1):132–164.
- Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnick, Boris Capitanu, Deren Kudeki, and J Stephen Downie. 2021. The gutenberghathitrust parallel corpus: A real-world dataset for noise investigation in uncorrected ocr texts.
- Athanasios Lagopoulos and Grigorios Tsoumakas. 2021. Self-citation analysis using sentence embeddings. *arXiv preprint arXiv:2105.05527*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Ernst Mayr. 1995. Darwin’s impact on modern thought. *Proceedings of the American Philosophical Society*, 139(4):317–325.
- Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. 2020. Neural machine translation with bert for post-ocr error detection and correction. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 333–336.
- Juri Opitz, Angel Daza, and Anette Frank. 2021. Weisfeiler-leman in the bamboo: Novel amr graph metrics and a benchmark for amr graph similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.
- Eszter Pal. 2014. Scientific societies in victorian england. *Review of Sociology*, 20:85–111.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Daniel N Rockmore, Chen Fang, Nicholas J Foti, Tom Ginsburg, and David C Krakauer. 2018. The cultural evolution of national constitutions. *Journal of the Association for Information Science and Technology*, 69(3):483–494.
- David A Smith, Ryan Cordell, and Abby Mullen. 2015. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, 27(3):E1–E15.
- Sandeep Soni, Lauren Klein, and Jacob Eisenstein. 2021. Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. *arXiv preprint arXiv:2103.07538*.
- Paul Vicinanza, Amir Goldberg, and Sameer B Srivastava. 2023. A deep-learning model of prescient ideas demonstrates that they emerge from the periphery. *PNAS nexus*, 2(1):pgac275.
- Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 349–357.

Evaluating Computational Representations of Character: An Austen Character Similarity Benchmark

Funing Yang and Carolyn Jane Anderson
Wellesley College
Wellesley, MA
carolyn.anderson@wellesley.edu

Abstract

Several systems have been developed to extract information about characters to aid computational analysis of English literature. We propose character similarity grouping as a holistic evaluation task for these pipelines. We present AustenAlike, a benchmark suite of character similarities in Jane Austen’s novels. Our benchmark draws on three notions of character similarity: a structurally defined notion of similarity; a socially defined notion of similarity; and an expert defined set extracted from literary criticism.

We use AustenAlike to evaluate character features extracted using two pipelines, BookNLP and FanfictionNLP. We build character representations from four kinds of features and compare them to the three AustenAlike benchmarks and to GPT-4 similarity rankings. We find that though computational representations capture some broad similarities based on shared social and narrative roles, the expert pairings in our third benchmark are challenging for all systems, highlighting the subtler aspects of similarity noted by human readers.

1 Introduction

There is growing interest in using computational techniques to analyze works of literary fiction. Several systems have been developed to automatically extract information about characters from English literary text (Bamman et al., 2014; Yoder et al., 2021). In this paper, we explore character similarity as a holistic evaluation task for literary pipelines. We use character similarity to explore the information about characters that is captured by the different kinds of features these pipelines extract: their events, utterances, and attributes.

Because characters can be similar along multiple axes, we construct a multi-part benchmark, AustenAlike, that uses three different notions of character similarity to group characters in Jane Austen’s novels. The first is a structurally defined

James Morland from *Northanger Abbey*

Sibling to heroine and single 20-year-old male clergy with income of £400/year

Social Pairings: Charles Hayter, Edward Ferrars, Robert Martin

Narrative Role Pairings: Isabella Knightley, John Dashwood, Margaret Dashwood, Susan Price, William Price, Elizabeth Elliot, Mary Musgrove, Jane Bennet, Mary Bennet, Kitty Bennet, Lydia Bennet

Expert Pairings: Edmund Bertram, Edward Ferrars, Henry Tilney, Philip Elton

Figure 1: Example character from AustenAlike

notion of similarity to group Austen’s characters: characters are similar if they fill similar narrative roles. The second is a socially defined notion of similarity: characters are similar if they share demographic features. The final benchmark takes a wisdom-of-the-crowd approach, but with an expert crowd: we extract comparisons of characters from four decades of *Persuasions*, a journal dedicated to the analysis of Austen’s work. Figure 1 shows an example of how these three views of character similarity can lead to different comparisons.

We use AustenAlike to explore how much information about characters is captured by the different kinds of features that literary pipelines extract. We extract character events, quotes, modifiers, and assertions using the BookNLP (Bamman et al., 2014; Sims et al., 2019) and FanfictionNLP Yoder et al. (2021) pipelines. We build character representations using contextualized embeddings of these features, and compare how well these representations align with the three sets of character groupings in the AustenAlike benchmarks. We also compare a non-feature-based approach by extracting similarity judgments from ChatGPT.

Our results show that event- and assertion-based representations capture more information about

character similarity than quote-based representations. Overall, however, we show that though computational representations capture some broad social and narratological similarities, there is a wide gap between the similarities they capture and the more nuanced similarities highlighted in our wisdom-of-the-expert-crowd benchmark. The best feature-based representations exhibit only medium correlations with expert rankings of character similarity, and GPT-4 lists the expert-identified most similar character in a top ten similarity list only half of the time. AustenAlike illustrates how much work remains to achieve nuanced computational representations of literary characters.

2 Related Work

There is a growing interest in applying computational methods to analyze literary fiction, both in analyses of large collections (*distant reading* (Moretti, 2013)) (Grayson et al., 2016; Jayanavar et al., 2015; Milli and Bamman, 2016) and of individual authors and works (Agarwal et al., 2013; Wang and Iyyer, 2019; Liebl and Burghardt, 2020). Though these projects range in scope, they share a foundation of feature extraction: literary evidence must be identified before it can be interpreted.

To facilitate computational analysis, a number of pipelines for extracting features from literary text have been developed (Bamman et al., 2014; Sims et al., 2019; Yoder et al., 2021; Ehrmantraut et al., 2023). In this paper, we focus specifically on features related to literary characters.

Character mentions The first step in computational studies of character is to identify character mentions using named entity recognition and coreference resolution. There is a large body of existing work on these tasks (Vala et al., 2015; Brooke et al., 2016; Roesiger and Teufel, 2014) given their complexity in a literary setting and their importance for downstream tasks.

Some pipelines further disambiguate character references in a *character clustering* step. BookNLP is a pipeline trained on data from LitBank, which provides annotated training data drawn from 19th- and early 20th-century English fiction, including annotations for named entity recognition (Bamman et al., 2019) and coreference resolution (Bamman et al., 2020). FanfictionNLP is a similar pipeline that is trained on and tailored to fanfiction.

Character features Once character mentions have been identified, the surrounding text can be used to extract information related to characters.

Some previous work focuses on character personality traits and emotions (Flekova and Gurevych, 2015). Kim and Klinger (2019) analyzes how emotions are expressed nonverbally in a corpus of fan fiction short stories, while Pizzolli and Strapparava (2019) train classifiers to identify personality traits in Shakespeare characters. The pipelines we study target more general descriptions: for FanfictionNLP, *assertions*, descriptions of physical and mental attributes; for BookNLP, modifiers and possessions.

What characters do and say is also of interest. Although quote attribution remains a challenging task with a number of approaches (He et al., 2013; Almeida et al., 2014; Muzny et al., 2017), it is useful for analyzing both the content and style of characters' speech (Dinu and Uban, 2017; Vishnubhotla et al., 2019). BookNLP extracts both events and quotes, while FanfictionNLP extracts only quotes.

There is also much work on mapping and analyzing relationships between characters (Elson et al., 2010; Lee and Yeung, 2012; Jayanavar et al., 2015; Agarwal et al., 2013; Wohlgenannt et al., 2016; Labatut and Bost, 2019). For instance, Chaturvedi et al. (2016) and Iyyer et al. (2016) automatically identify how relationships between characters change over the course of narratives.

Character models Once character features are extracted, they can be used to build computational representations of characters. Some work seeks to classify characters into types (Chambers and Jurafsky, 2009; Valls-Vargas et al., 2021; Stammbach et al., 2022; Bamman et al., 2014). For instance, Jahan and Finlayson (2019) propose a narratologically-grounded framework for character identification and a simple rule-based system for extracting characters and their roles.

Others explore authorial decisions in representing characters (Bullard and Ovesdotter Alm, 2014) or how they evolve over retellings (Besnier, 2020).

Some approaches learn character representations directly. Grayson et al. (2016) show that word embeddings learned from 19th-century works of fiction provide insight into characters. Holgate and Erk (2021) learn vector representations using masked entity prediction as a training objective. Most similar to our work, Inoue et al. (2022) propose a benchmark for evaluating character representations.

Their work takes a broad multi-author, multi-task perspective, while ours dives more deeply into characters by a single author, exploring character similarity from three different angles.

3 A Three-Part Benchmark for Evaluating Character Similarity

Character similarity is a multi-faceted concept. Two characters may play the same role in a narrative or follow the same plot trajectory. They may have similar personality traits or fill similar social roles. AustenAlike uses a multi-faceted approach to character similarity that explores three aspects of literary characterhood: shared narrative roles, shared social characteristics, and pairwise comparisons from expert analysis.¹ The AustenAlike benchmark focuses on characters from the six Jane Austen novels published within or immediately after her lifetime: *Sense and Sensibility*, *Pride and Prejudice*, *Mansfield Park*, *Emma*, *Persuasion*, and *Northanger Abbey*. We include all named characters who speak more than once, except those who die in the first chapter.

3.1 Social Characteristics

Jane Austen’s novels highlight how her character’s choices are impacted by their position in society. Although her characters struggle to varying degrees to reconcile their desires with constraints imposed by gender, rank, and wealth, these social characteristics play a large part in determining the options available to them within the novel.

We consider five demographic dimensions that define social relationships within Austen’s writing: marital status, gender, rank, age, and wealth. There are other social characteristics that demarcated opportunities within Austen’s historical context, such as race and nationality; however, the characters under consideration are homogeneously White and English.² A summary of the social categories and the size of each group is in Appendix A.

Rank Although almost all of Jane Austen’s characters belong to the upper middle or lower upper classes, their relative social rank is nonetheless important to their prospects. Most characters are gentry: independently wealthy, often landowners. Lower-ranked characters belong to professions.

¹The dataset and support code are available at <https://github.com/Wellesley-EASEL-lab/AustenAlike>.

²Given the exclusion of Austen’s unfinished *Sanditon*.

Following social conventions of the time, an unmarried woman has her father’s rank and a married woman her husband’s.

Wealth Austen novels center on questions of wealth, particularly as they relate to marital prospects. As a result, the wealth of unmarried characters is typically stated. The wealth of married characters is not always stated. We draw on estimates from Heldman (1990) and Toran (2015).

Gender The genders of all Austen characters are overt and stable. All characters are Male or Female.

Age Character ages are reasonably stable as almost all plot events take place within a year. If a character’s age is not mentioned, we estimate from the ages of their family members.

Marital status Marital status is a key social characteristic of Austen characters. We divide characters into four groups: Single, Married, Widowed, and Transitional, a group comprising the handful of characters whose marital status changes before the climax of the novel.

3.2 Narrative Roles

Another way in which characters can resemble each other is in the role they play in the narrative structure of the work. We define seven narrative roles:

- Heroine: each novel has at least one protagonist who is an unmarried woman seeking a marriage partner.
- Hero: the character that each protagonist marries at the novel’s end.
- Deceiver: each novel features a character who sets key events in motion by lying about himself or the heroine.
- Rival: an alternate love interest for the hero.
- Wooer: an alternate love interest for the heroine.
- Parents: the parents of the heroine.
- Siblings: the siblings of the heroine.

These groupings are shown in Table 1.

3.3 Wisdom-of-the-Experts Character Pairs

In our most fine-grained benchmark, we look at characters who have been identified as similar by literary scholars. We use a wisdom-of-the-crowds approach, but with an expert crowd: authors of

Heroines:	Emma Woodhouse, Elizabeth Bennet, Elinor Dashwood, Marianne Dashwood, Fanny Price, Catherine Morland, Anne Elliot
Heroes:	George Knightley, Fitzwilliam Darcy, Edward Ferrars, Edmund Bertram, Henry Tilney, Frederick Wentworth, Colonel Brandon
Deceivers:	John Thorpe, George Wickham, John Willoughby, William Elliott, Henry Crawford, Frank Churchill
Rivals:	Caroline Bingley, Lucy Steele, Louisa Musgrove, Mary Crawford, Harriet Smith
Woosers:	Henry Crawford, William Elliot, Philip Elton, Charles Musgrove, William Collins, John Thorpe
Siblings:	Marianne Dashwood, Jane Bennet, Lydia Bennet, Mary Bennet, Kitty Bennet, Susan Price, Mary Musgrove, Elizabeth Elliot, Isabella Knightley, James Morland, William Price
Parents:	Mr. Bennet, Sir Walter Elliot, Lieutenant Price, Mr. Woodhouse, Mrs. Bennet, Mrs. Dashwood, Mrs. Price, Mrs. Morland

Table 1: Narrative Roles benchmark summary

articles published in *Persuasions*, the Jane Austen Society of North America’s peer-reviewed journal.

We manually reviewed 43 volumes of *Persuasions* to create a set of character pairings. We extract all instances of a similarity or shared property discussed in an article. When an article mentions a similarity between more than two characters, we add all pairings from the set. The resulting dataset contains 5740 character comparison pairs.

The identified comparisons are diverse, encompassing traits from our other benchmarks, such as rank, age, and narrative role, as well as more nuanced commonalities. For instance, *Persuasions* authors describe Edward Ferrars and Frank Churchill as similar because both are secretly engaged; Emma Woodhouse and Lady Catherine de Bourgh because they oversee charitable work; and Isabella Thorpe and Lydia Bennet because of their flirtatiousness. These expert-identified pairings provide a comprehensive view of character similarity.

4 Building Computational Representations of Character

We build computational representations of character from the output of two literary pipelines. We construct representations out of the features they extract: for BookNLP, events, quotes, and modifiers; for FanfictionNLP, quotes and assertions.

4.1 Character Mentions

We use each pipeline to identify all character mentions, perform coreference resolution, and aggregate character mentions. We then merge and filter character clusters using a handwritten alias map for

Austen character names.

4.2 Feature Embeddings

We retrieve contextualized embeddings for each kind of feature. For events and modifiers, which are single words, we retrieve a contextualized embedding of the word in its context using T5 (11B) (Raffel et al., 2020). For quotes and assertions, we retrieve sentence embeddings using NV-Embed (7.85B) (Lee et al., 2024). We center each kind of feature embedding by subtracting the mean of all embeddings for the feature.

For each feature and character, we construct a character representation by averaging the embeddings of the character’s features. For events, we average the character’s agent events and patient events separately and concatenate the vectors. This process produces 5 representations per character: an assertion vector, a modifier vector, an event vector, and two quote vectors (one per pipeline).

Having produced these 5 representations for each character, we are interested in exploring the effectiveness of each kind of feature-based representation in capturing character similarity. Thus, we compute each result presented in Section 6 for each of the 5 representations.

4.3 GPT-4 comparison

We provide a non-featured-based comparison by querying a pretrained large language model, GPT-4 (Achiam et al., 2023), for character similarity rankings. Given the popularity of Austen’s work, we assume that GPT-4’s training data contains all six novels and many web pages discussing them.

We extract character similarities using three approaches: asking GPT-4 to select the most similar character from a list of all benchmark characters; asking GPT-4 to select the most similar character and explain its choice; and asking GPT-4 to choose the ten most similar characters from a list of all benchmark characters. We repeat each experiment 5 times (further details in Appendix B).

5 Evaluating character similarity

We have proposed three benchmarks that capture different aspects of character similarity. For the social and narrative roles benchmarks, we are interested in the similarity between characters in the same groupings. For the expert benchmark, we are interested in whether characters are most similar to those they are paired with by experts.

5.1 Grouping evaluation

The Social and Narrative benchmarks define groupings of characters. We explore how strongly these groupings are captured by computational character representations using two evaluation metrics.

In-group Cosine Similarity We explore whether characters are more similar to characters within their group than those outside of their group. We compute the average cosine similarity between a grouped character and all other group members, and compare it to the average cosine similarity between the character and non-group characters. We call this *in/out-group cosine similarity difference*.

Most Similar Character We also ask whether very similar characters come from the same groups. We count how often the single character with highest cosine similarity to the target character belongs to the same group.

5.2 Pairing evaluation

For the Expert benchmark, we measure the extent to which the cosine similarities of each kind of representation align with the expert-identified pairs using three metrics:

Correlation We look at the correlation between cosine similarity of two character representations and the number of times experts describe the two characters as similar. We calculate Pearson’s ρ to measure the strength of the correlation.

Ranking similarity Literary experts may be more interested in identifying highly similar characters than in quantifying degrees of dissimilarity. We identify the ten most similar characters according to experts and to cosine similarity, and compute the alignment between the lists using Jaccard similarity. Jaccard similarity measures the intersection of the groups divided by their union. If the two lists are completely different, their Jaccard similarity is 0; if they mostly agree, it is close to 1.

Top character in ten-most similar Finally, we focus on the top expert-identified pairings. We count how often the character who experts pair most with a target character has one of the ten highest cosine similarities to the target character.

6 Results

We explore how well computational representations of character capture aspects of character similarity using the three-part AustenAlike benchmark.

6.1 Narrative Roles Benchmark

The narrative roles benchmark explores similarity between characters who play similar roles in the plot of a novel. Are heroines similar to other heroines? Are parents similar to other parents? If parents are described similarly to other parents, assertion- and modifier-based representations should capture their similarity; if they say and do similar things as other parents, their quote- and event-based representations should be similar.

6.1.1 Are same-role characters more similar?

We test whether characters who share the same narrative role are more similar than characters who do not. We compare the average cosine similarity of representations within a narrative role group to their similarity to non-group members. We compute the in-group and out-group scores for each character in a target role group and average them.

Figure 2 plots the cosine similarity for characters within the same narrative role group compared to characters outside of the group. We observe that event- and assertion-based representations are the best at showing dissimilarity for characters outside of the role group. The FanfictionNLP quote-based representations show the weakest differences between in-group and out-group members.

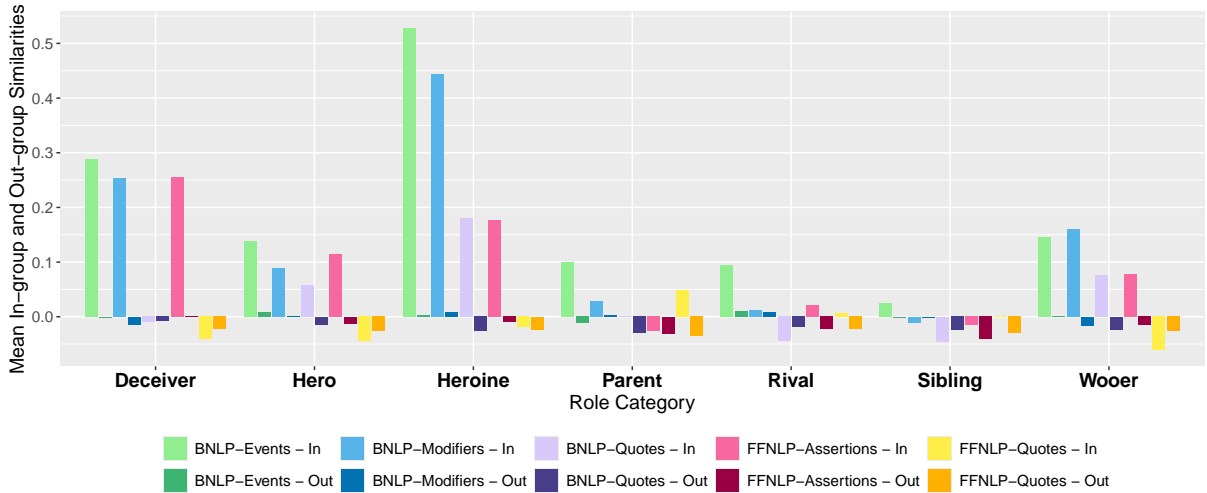


Figure 2: Narrative Role Benchmark: Mean cosine similarities between same-group characters and other characters by representation type.

System	Hero	Heroine	Deceiver	Rival	Wooer	Parent	Sibling
FanfictionNLP Assertions	0.29	0.43	0.33	0	0	0.18	0.29
BookNLP Events	0	1	0.36	0.09	0.18	0.35	0
BookNLP Modifiers	0	0.86	0.33	0.2	0	0.27	0.18
BookNLP Quotes	0.13	0.78	0.57	0.33	0.43	0.08	0
FanfictionNLP Quotes	0	0.43	0	0.14	0	0.18	0.08
GPT-4	0.43	0.43	0.5	0	0	0.33	0.25
GPT-4 Reasoning	0.86	1	0.83	0.17	0.5	0.42	0.08

Table 2: Narrative Role Benchmark: Average occurrence of most similar character in same narrative role group by character representation. Characters from same novel are excluded.

6.1.2 Is the most similar character from the same group?

We also explore whether a target character’s most similar character belongs to the same narrative role group. For each character, we count how often the character with highest cosine similarity belongs to the same role group. Feature-based representations can be skewed towards same-novel similarity: for instance, characters in *Northanger Abbey* are more likely to engage in reading events since this is a theme of the novel. We therefore explore results with and without characters from the same novel.

Table 2 reports how often the most similar character occurs in the same role group, with same-novel characters excluded (inclusive version in Appendix C). We see marked differences between categories. Heroines are frequently similar to heroines for all representations, while other groups have lower rates of same-group membership.

The BookNLP quote representations capture narrative role similarity better than the FanfictionNLP quote representations, perhaps because BookNLP

is trained on literary fiction. However, FanfictionNLP assertions perform competitively in two of the most challenging categories for feature-based representations, Hero and Sibling.

We observe that GPT-4, when asked to justify its decision, is more sensitive to narrative role than the feature-based representations in about half of the categories. However, without reasoning-prompting, it is no better than the feature-based representations, identifying selecting a heroine as the most similar to heroines only 43% of the time.

Qualitatively, a challenging aspect of this benchmark seems to stem from young single characters with different narrative roles. Like heroes and heroines, deceivers, wooers, and rivals tend to be unmarried and of a similar age. We observe that heroes tend to be similar to deceivers (10/69 out-group cases) and vice versa (12/50 out-group cases), and rivals to heroines (26/64) and vice versa (6/31 out-group cases), aligning with the social characteristics of each set. The error patterns for the remaining categories seem less clear, perhaps reflecting the

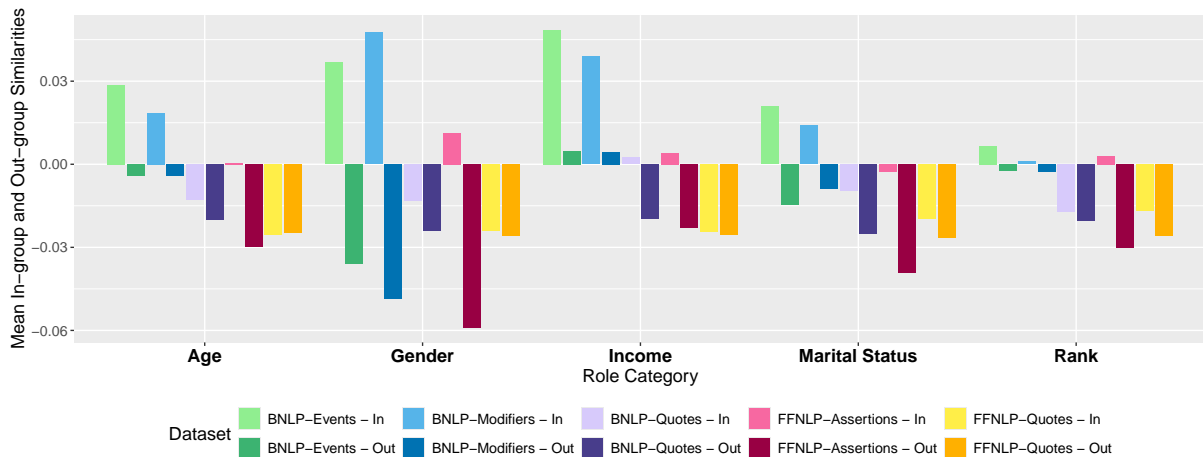


Figure 3: Social Benchmark: average differences in cosine similarity between same-group characters and other characters by character representation and social role group.

System	Age	Gender	Income	Marital Status	Rank
FanfictionNLP Assertions	0.16	0.9	0.02	0.5	0.34
BookNLP Events	0.23	0.76	0.07	0.51	0.29
BookNLP Modifiers	0.22	0.80	0.05	0.46	0.19
BookNLP Quotes	0.06	0.63	0.15	0.42	0.26
FanfictionNLP Quotes	0.13	0.54	0.02	0.3	0.25
GPT-4	0.26	0.80	0.21	0.52	0.42
GPT-4 Reasoning	0.32	0.98	0.07	0.58	0.39

Table 3: Social Benchmark: average occurrence of most similar characters in the same social group by character representation. Characters from same novel are excluded.

limited mentions of parent characters and the more heterogeneous characteristics of siblings.

6.2 Social Benchmark

The second AustenAlike benchmark evaluates character similarity on the basis of social characteristics. It groups characters based on five demographic features: rank, wealth, gender, age, and marital status. Modifiers and assertions may directly describe these characters. However, given that a character’s social status delimits the set of actions and utterances available to them, we also expect event- and quote-based representations to echo back similarities based on these characteristics.

6.2.1 How similar are characters with shared social characteristics?

We explore whether characters within the same group in each of the social categories are most similar to each other. Figure 3 plots the average cosine similarity for characters within the same social group compared with non-group members.

We observe that the event-based representations

are the most reliable for distinguishing social similarity. Gender shows the sharpest in-group/out-group differences for all three categories, followed by income. Quote-based representations struggle to capture similarity by social group: the FanfictionNLP quote-based representations do not capture differences for any of the criteria, while the BookNLP quote-based representations show only a (weak) in-group/out-group difference for income.

6.2.2 Is the most similar character from the same group?

We also focus more narrowly on the top-most similar character. Table 3 shows how often the character with the highest cosine similarity to the target character occurs in the same social group. Top character representations most commonly share gender and then marital status. This makes sense, since Austen’s plots center around courtship: these key aspects of identity should be reflected in how they are described and the events they participate in.

GPT-4’s similarity judgments align with social characteristics more strongly than any of the

Dataset	Pearson’s ρ	Jaccard Similarity	Top in Top 10
FanfictionNLP Assertions	0.29	0.03	0.69
BookNLP Events	0.4	0.02	0.34
BookNLP Modifiers	0.28	0.01	0.29
BookNLP Quotes	0.27	0.03	0.56
FanfictionNLP Quotes	0.15	0.02	0.49
GPT-4	-	-	0.52
GPT-4 Reasoning	-	-	0.56
GPT-4 Top Ten List	-	0.02	-

Table 4: Expert Benchmark: measures of alignment between expert pairing counts and computational similarity.

feature-based representations. Quote-based representations do not seem to capture similarity by social characteristics as well as the other feature-based representations in most categories.

6.3 Expert Benchmark

Our last benchmark takes an expert wisdom-of-the-crowd approach. The expert benchmark contains counts of character similarity pairings. We compare these pairing counts to the cosine similarity between the computational representations of the two characters to evaluate how well computational representations aligns with expert judgments of character similarity.

6.3.1 Does cosine similarity correlate with expert judgments?

We examine how well computational character representations align with expert judgments by measuring the correlation between expert character pairings and cosine similarity. We posit that high quality computational representations should produce higher cosine similarity between the characters that are more frequently deemed similar by experts.

Table 4 shows the correlation between expert pairing counts and cosine similarity for each of the computational representations.

Overall, we observe moderate positive correlations between the cosine similarity of character representations and the number of expert similarity pairings. The BookNLP event representations correlate most strongly with expert pairings, while the FanfictionNLP quote-based representations correlate less strongly than other feature-based representations. This converges with our social and narratological similarity findings.

Although the expert benchmark is useful in differentiating among feature-based representations, it is also important to note that none of the feature-based representations are strongly correlated with

expert judgments. This shows that there are many aspects of character similarity that are apparent to human readers that remain uncaptured in the computational character representations we explore.

6.3.2 Is there agreement on the most similar characters?

Correlations between cosine similarity and expert pairing counts may be skewed by very dissimilar characters, whose expert pairings are few. We also look at two measures of agreement for the most similar characters.

For each character, we retrieve the ten characters with the highest cosine similarity, and the ten characters with whom they are most frequently paired by experts. We then measure agreement by computing the Jaccard similarity of the two sets.

Table 4 shows the average Jaccard similarity these top ten sets. The Jaccard scores are uniformly low, indicating that cosine similarity tends not to identify the same set of highly similar characters as experts. Interestingly, GPT-4 does not appear any more successful at identifying expert-aligned similar characters than the feature-based approaches, despite its success in identifying socially and narratologically similar characters.

We also examine how often the single character that experts compare most to a target character occurs within the target’s top ten closest representations by cosine similarity. Table 4 shows the average success on this lenient measure.

Even with this easier measure, the expert benchmark is quite challenging. GPT-4 includes the expert top character in its top ten list only half of the time. The best feature-based representation, FanfictionNLP assertions, include it 69% of the time. Since this is a very lenient measure of success, this illustrates the large gaps that remain between similarity by computational representations of character, pretrained LLM understanding of character

similarity, and expert evaluations.

7 Conclusion

We present AustenAlike, a three-part Jane Austen benchmark for evaluating multiple aspects of character similarity: narrative role similarity, social similarity, and expert judgments of character similarity drawn from prior scholarly analysis. We use AustenAlike to evaluate five computational representations of character built atop features extracted by pipelines for analyzing English literature.

We find that event- and assertion-based representations tend to capture character similarity better than quote-based representations. Overall, however, our results show how much work still remains to be done to improve computational representations of character: feature-based representations and GPT-4 alike struggle to place the expert-identified most similar character in their top ten lists of character similarity. We hope that by providing a multi-faceted benchmark with expert judgments, AustenAlike can guide future work on computational representations of character.

Limitations

We have evaluated five kinds of feature-based character representations across two systems. However, our approach has a number of limitations.

Noisy Character Data Both pipelines produce character clusters with some amount of inconsistency and error. In some cases, the pipelines failed to resolve multiple ways of referring to the same character (*Miss Tilney*, *Eleanor Tilney*). We post-process the output with an Austen-specific alias map; to extend our work to other works of literature, this post-processing step would need to be manually extended.

Missing Characters Both pipelines failed to extract features for some characters included in our benchmark. BookNLP failed to identify twelve characters and FanfictionNLP failed to identify four. This was most impactful in the siblings and parents subsets of the narrative roles benchmark.

Generalizability Our benchmark focuses on characters from the work of Jane Austen. As a result, it may favor methods of deriving computational representations that are trained on similar literary text. This may affect our comparison of FanfictionNLP and BookNLP quotes, as noted above.

Combining Character Data In this paper, we compare 5 different kinds of feature-based representations: events, assertions, modifiers, and quotations extracted from two pipelines. However, it would also be possible to combine these different sources of information about a character, and use them together. Future work could explore this kind of merged representation.

Ethics Statement

Our work does not involve any human data. The literary works we analyze are in the public domain. The computational resources involved in our experiments are also modest: all contextualized embeddings were extracted using less than 12 hours on a single Nvidia RTX A6000 GPU.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. [Automatic extraction of social networks from literary text: A case study on alice in wonderland](#). In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1202–1208, Nagoya, Japan. Asian Federation of Natural Language Processing.
- Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. [A joint model for quotation attribution and coreference resolution](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg, Sweden. Association for Computational Linguistics.
- David Bamman, Olivia Lewke, and Anya Mansoor. 2020. [An annotated dataset of coreference in English literature](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.
- David Bamman, Sejal Papat, and Sheng Shen. 2019. [An annotated dataset of literary entities](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.
- David Bamman, Ted Underwood, and Noah A. Smith. 2014. [A Bayesian mixed effects model of literary character](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

- (*Volume 1: Long Papers*), pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.
- Clément Besnier. 2020. [History to myths: Social network analysis for comparison of stories over time](#). In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9, Online. International Committee on Computational Linguistics.
- Julian Brooke, Adam Hammond, and Timothy Baldwin. 2016. [Bootstrapped text-level named entity recognition for literature](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Joseph Bullard and Cecilia Ovesdotter Alm. 2014. [Computational analysis to explore authors’ depiction of characters](#). In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 11–16, Gothenburg, Sweden. Association for Computational Linguistics.
- Nathanael Chambers and Dan Jurafsky. 2009. [Unsupervised learning of narrative schemas and their participants](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.
- Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. [Modeling evolving relationships between characters in literary novels](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).
- Liviu P. Dinu and Ana Sabina Uban. 2017. [Finding a character’s voice: Stylome classification on literary characters](#). In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 78–82, Vancouver, Canada. Association for Computational Linguistics.
- Anton Ehrmanntraut, Leonard Konle, and Fotis Jannidis. 2023. [LLpro: A Literary Language Processing Pipeline for German Narrative Texts](#). In *Conference on Natural Language Processing*.
- David Elson, Nicholas Dames, and Kathleen McKeown. 2010. [Extracting social networks from literary fiction](#). In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.
- Lucie Flekova and Iryna Gurevych. 2015. [Personality profiling of fictional characters using sense-level links between lexical resources](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816, Lisbon, Portugal. Association for Computational Linguistics.
- Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. 2016. [Novel2vec: Characterising 19th century fiction via word embeddings](#). In *Irish Conference on Artificial Intelligence and Cognitive Science*.
- Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. [Identification of speakers in novels](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.
- James Heldman. 1990. How wealthy is Mr. Darcy – Really? Pounds and Dollars in the World of *Pride and Prejudice*. *Persuasions*, 12:38–49.
- Eric Holgate and Katrin Erk. 2021. [“politeness, you simpleton!” retorted \[MASK\]: Masked prediction of literary characters](#). In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 202–211, Groningen, The Netherlands (online). Association for Computational Linguistics.
- Naoya Inoue, Charuta Pethe, Allen Kim, and Steven Skiena. 2022. [Learning and evaluating character representations in novels](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1008–1019, Dublin, Ireland. Association for Computational Linguistics.
- Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. [Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.
- Labiba Jahan and Mark Finlayson. 2019. [Character identification refined: A proposal](#). In *Proceedings of the First Workshop on Narrative Understanding*, pages 12–18, Minneapolis, Minnesota. Association for Computational Linguistics.
- Prashant Jayannavar, Apoorv Agarwal, Melody Ju, and Owen Rambow. 2015. [Validating literary theories using automatic social network extraction](#). In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 32–41, Denver, Colorado, USA. Association for Computational Linguistics.
- Evgeny Kim and Roman Klinger. 2019. [Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics.

- Vincent Labatut and Xavier Bost. 2019. [Extraction and analysis of fictional character networks: A survey](#). *ACM Comput. Surv.*, 52(5).
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. [Nv-embed: Improved techniques for training llms as generalist embedding models](#).
- John Lee and Chak Yan Yeung. 2012. [Extracting networks of people and places from literary texts](#). In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 209–218, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.
- Bernhard Liebl and Manuel Burghardt. 2020. [“shakespeare in the vectorian age” – an evaluation of different word embeddings and NLP parameters for the detection of shakespeare quotes](#). In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 58–68, Online. International Committee on Computational Linguistics.
- Smitha Milli and David Bamman. 2016. [Beyond canonical texts: A computational analysis of fanfiction](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2048–2053, Austin, Texas. Association for Computational Linguistics.
- Franco Moretti. 2013. *Distant Reading*. Verso, London.
- Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. [A two-stage sieve approach for quote attribution](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.
- Daniele Pizzolli and Carlo Strapparava. 2019. [Personality traits recognition in literary texts](#). In *Proceedings of the Second Workshop on Storytelling*, pages 107–111, Florence, Italy. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Ina Roesiger and Simone Teufel. 2014. [Resolving coreferent and associative noun phrases in scientific text](#). In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–55, Gothenburg, Sweden. Association for Computational Linguistics.
- Matthew Sims, Jong Ho Park, and David Bamman. 2019. [Literary event detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.
- Dominik Stambach, Maria Antoniak, and Elliott Ash. 2022. [Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data](#). In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.
- Katherine Toran. 2015. The economics of jane austen’s world. *Persuasions On-Line*, 36:1817–1853.
- Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. [Mr. benet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.
- Josep Valls-Vargas, Santiago Ontañón, and Jichen Zhu. 2021. [Toward character role assignment for natural language stories](#). *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.
- Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2019. [Are fictional voices distinguishable? classifying character voices in modern drama](#). In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–34, Minneapolis, USA. Association for Computational Linguistics.
- Shufan Wang and Mohit Iyyer. 2019. [Casting Light on Invisible Cities: Computationally Engaging with Literary Criticism](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1291–1297, Minneapolis, Minnesota. Association for Computational Linguistics.
- Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. 2016. [Extracting social networks from literary text with word embedding tools](#). In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 18–25, Osaka, Japan. The COLING 2016 Organizing Committee.
- Michael Yoder, Sopan Khosla, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, and Carolyn Rosé. 2021. [FanfictionNLP: A text processing pipeline for fanfiction](#). In *Proceedings of the Third Workshop on Narrative Understanding*, pages 13–23, Virtual. Association for Computational Linguistics.

Category	Group	N
Rank	Nobility	2
	Titled Gentry	15
	Gentle	48
	New Gentle	5
	Clergy	12
	Military	13
	Profession	14
Wealth	£50	8
	£51-£250	7
	£251-£500	9
	£501-£1000	8
	£1001-£3000	6
	£3001+	5
Gender	Male	50
	Female	59
Age	< 18	8
	18-20	13
	21-24	16
	25-27	18
	28-30	12
	31-40	13
	41-50	19
	51+	10
Marital Status	Single	48
	Transitional	6
	Married	42
	Widowed	13

Table 5: Social Characteristics benchmark summary

A Further Details of Benchmark Construction

A.1 Social Benchmark

Rank To achieve a more even balance across groups, we partition untitled gentry into two groups: New Gentle, characters whose fathers were not gentlemen, and Gentle, representing more established gentry. We consolidate professional characters into three groups: a military group encompassing the army and navy; a professional group encompassing business, law, and farming; and a clergy group. This totals six categories: New Gentle, Gentle, Gentry, Military, Profession, Clergy, and Nobility.

Wealth Wealth for women is generally reported as a total sum, while men’s fortunes are typically stated in terms of yearly income. We convert all figures to yearly incomes assuming the 5% yearly dividend standard during Austen’s time (Toran, 2015).

Marital Status Marital status tends to remain stable until the end of each novel: although many single characters marry, most marriages take place in the last chapter.

A.2 Narrative Roles Benchmark

Heroines All Jane Austen novels involve young people finding marriage partners. Each novel has at least one protagonist who is an unmarried woman seeking a marriage partner. *Sense and Sensibility* focuses on a pair of sisters who both marry by the end of the novel; we treat both as protagonists/heroines. Heroines should be particularly easy to distinguish from other narrative roles since they are the main viewpoint characters in Austen’s novels.

Heroes We use the term *hero* for the character that each protagonist marries at the novel’s end.

Deceiver Each of Austen’s novels features at least one character who lies in a way that sets key events in motion. Frequently, this character misrepresents himself to the heroine in a key way (Wickham in *Pride and Prejudice*; Willoughby in *Sense and Sensibility*); in other cases, the character lies to conceal an ulterior motive (William Elliot in *Persuasion*; Frank Churchill in *Emma*). In one case, this character spreads lies about the heroine herself (John Thorpe in *Northanger Abbey*).

Rivals and Wooers In each of the six novels, there is at least one character who serves as a rival, an alternate love interest for the hero. In all but one novel (*Sense & Sensibility*), there is a character who unsuccessfully courts the heroine; we refer to these characters as *wooers*.

Family roles Austen’s novels are concerned with domestic settings and interactions within a relatively confined society. As a result, there are numerous family members. We look at two groups: parents and siblings. In the case of *Mansfield Park*, in which the heroine is raised in her uncle’s family, we considered including her guardians but excluded them to be consistent with other mentors (Lady Russell in *Persuasion*) and temporary guardians (the Allens in *Northanger Abbey*).

B Further Details of GPT-4 Experiments

We run three experiments to extract character similarities from GPT-4: a top character experiment, a top character experiment with reasoning, and a top

ten characters experiment. We run each experiment five times at temperature=0.2.

The prompts are shown below (full list of characters omitted for readability). c represents the name of the target character, and $cIndex$ is that character’s number in the list.

Top Character Prompt

Consider the following list of Jane Austen characters:

1. Anna Weston
2. Augusta Elton
- ...
108. Sir John Middleton
109. Thomas Palmer

Which character is c most similar to (other than c)? Respond with only a number. Do not choose $cIndex$.

Top Character with Reasoning Prompt

Consider the following list of Jane Austen characters:

1. Anna Weston
2. Augusta Elton
- ...
108. Sir John Middleton
109. Thomas Palmer

Which character is c most similar to (other than c)? Describe your reasoning and then reply with the number of the character. Do not choose $cIndex$.

Top Ten Characters Prompt

Consider the following list of Jane Austen characters:

1. Anna Weston
2. Augusta Elton
- ...
108. Sir John Middleton
109. Thomas Palmer

List the 10 characters that are most similar to c (other than c). Consider characters from all Austen novels. Reply with just their numbers. Do not choose $cIndex$.

C Further Results

C.1 Narrative Role Benchmark

Table 6 shows how often the most similar character is within the same narrative role set as the target

character, with all books included. Table 2 excludes characters from the same book.

C.2 Social Benchmark

Table 7 shows how often the most similar character is within the same social role set as the target character, with all books included. Table 3 excludes characters from the same book.

C.3 Expert Benchmark

Tables 8 and 9 shows Pearson’s ρ correlations between cosine similarity and expert pairing counts by novel, with characters from the same novel included and excluded respectively.

System	Hero	Heroine	Deceiver	Rival	Wooser	Parent	Sibling
FanfictionNLP Assertions	0.14	0.36	0.17	0	0	0.18	0.25
BookNLP Events	0.07	1	0.33	0.08	0.17	0.36	0
BookNLP Modifiers	0	0.86	0.33	0.25	0	0.27	0.18
BookNLP Quotes	0.07	0.64	0.33	0.17	0.25	0.09	0
FanfictionNLP Quotes	0.14	0.21	0	0.08	0	0.14	0.08
GPT-4	0.43	0.43	0.5	0	0	0.33	0.25
GPT-4 Reasoning	0.86	1	0.83	0.17	0.5	0.42	0.08

Table 6: Narrative Role Benchmark: Average occurrence of most similar character in same narrative role group by character representation. Characters from same novel are included.

System	Age	Gender	Income	Marital Status	Rank
FanfictionNLP Assertions	0.18	0.75	0.13	0.52	0.41
BookNLP Events	0.23	0.77	0.13	0.51	0.30
BookNLP Modifiers	0.21	0.78	0.07	0.46	0.19
BookNLP Quotes	0.09	0.58	0.15	0.40	0.34
FanfictionNLP Quotes	0.10	0.49	0.05	0.37	0.34
GPT-4	0.26	0.80	0.21	0.52	0.42
GPT-4 Reasoning	0.32	0.98	0.07	0.58	0.39

Table 7: Social Benchmark: average occurrence of most similar characters in the same social group by character representation. Characters from same novel are included.

Novel	<i>Emma</i>	<i>MP</i>	<i>NA</i>	<i>Pers.</i>	<i>P&P</i>	<i>S&S</i>	All
FanfictionNLP Assertions	0.30	0.38	0.29	0.27	0.23	0.28	0.29
BookNLP Events	0.44	0.44	0.43	0.31	0.37	0.43	0.4
BookNLP Modifiers	0.31	0.31	0.25	0.26	0.26	0.29	0.28
BookNLP Quotes	0.26	0.28	0.28	0.20	0.24	0.35	0.27
FanfictionNLP Quotes	0.21	0.20	0.15	0.11	0.10	0.11	0.15

Table 8: Expert Benchmark: Pearson’s ρ correlation between cosine similarity and expert pairing count by character representation. Character pairs with no expert mentions are excluded.

Novel	<i>Emma</i>	<i>MP</i>	<i>NA</i>	<i>Pers.</i>	<i>P&P</i>	<i>S&S</i>	All
FanfictionNLP Assertions	0.3	0.38	0.34	0.23	0.33	0.27	
BookNLP Events	0.47	0.48	0.50	0.47	0.45	0.48	
BookNLP Modifiers	0.34	0.37	0.33	0.38	0.34	0.35	
BookNLP Quotes	0.16	0.11	0.27	0.13	0.28	0.27	
FanfictionNLP Quotes	-0.01	0.04	0.15	0.04	0.04	-0.07	

Table 9: Expert Benchmark: Pearson’s ρ correlation between cosine similarity and expert pairing count by character representation. Characters from the same novel are excluded. Character pairs with no expert mentions are excluded.

Investigating Expert-in-the-Loop LLM Discourse Patterns for Ancient Intertextual Analysis

Ray Umphrey[†]

University of the Cumberland
ray.umphrey@ucumberland.edu

Jesse Roberts[†]

Tennessee Tech University
jtroberts@tntech.edu

Lindsey Roberts

Tennessee Tech University

Abstract

This study explores the potential of large language models (LLMs) for identifying and examining intertextual relationships within biblical, Koine Greek texts. By evaluating the performance of LLMs on various intertextuality scenarios the study demonstrates that these models can detect direct quotations, allusions, and echoes between texts. The LLM’s ability to generate novel intertextual observations and connections highlights its potential to uncover new insights. However, the model also struggles with long query passages and the inclusion of false intertextual dependences, emphasizing the importance of expert evaluation. The expert-in-the-loop methodology presented offers a scalable approach for intertextual research into the complex web of intertextuality within and beyond the biblical corpus.

1 Introduction

Intertextuality, coined by Julia Kristeva (Kristeva, 1980), proposes that the meaning which should be understood as intended by an author is that most relevant in the common zeitgeist. However, like stacked layers of a fuzzy neural network (Kwan and Cai, 1994), authors’ work transforms the meaning of the input linguistic “sign” so that the understanding incumbent on future works is forever changed.

As shown in Figure 1, many contemporary references to Satan depend more (but not exclusively) upon Milton than the Bible, while Genesis provides the base understanding invoked and transformed by Milton (Allen, 2011). This notion of an intertextual network is an extension of the intertextual graphs discussed in Kuznetsov et al. (2022) as it supposes a non-linear transformation is applied by the author in the work to the inherited linguistic sign.

When the date of authorship and lineage of a text are known, the structure of the network can be

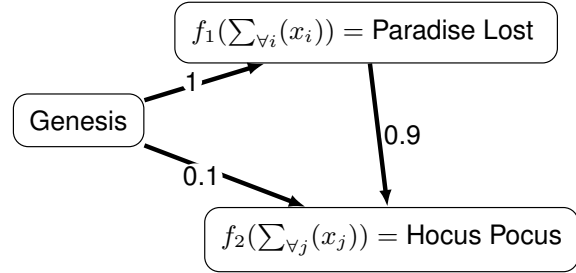


Figure 1: Example approximated intertextual dependence over Genesis, Paradise Lost (Milton, 2005), and Hocus Pocus (Ortega, 1993). Future works depend on past related texts with some weight. Subsequent work transforms the combined inherited weighted representations like a neural activation function.

asserted and the weights may be studied to understand intertextualities. However, in biblical studies, texts are often dated imprecisely or may not have been sufficiently distributed to support confident assertions of textual dependence. Therefore, the presence of a strong intertextual dependence can provide both important situational and interpretational guidance for the passage.

While many such intertextual dependencies have been documented, the relative youth of the interpretational framework and the vast number of relatively co-temporal texts makes searching for potential intertextualities tedious.

Table 1: List of Contributions

Contribution 1	<i>We find that the query length has a significant impact while long corpora pose no issue to retrieval.</i>
Contribution 2	<i>We identify a pattern of LLM usage that augments an expert’s ability to catalogue and evaluate the weight of intertextual dependencies in Koine Greek.</i>
Contribution 3	<i>We identify types of errors made by the LLM.</i>
Contribution 4	<i>We show that LLMs perform intertextual analysis by retrieving known intertextualities and by directly comparing the similarity of passages.</i>
Contribution 5	<i>We identify a strong previously undocumented intertextuality through LLM support.</i>

This paper seeks to use large language models (LLMs) to identify and examine the strength of intertextual relationships within biblical texts. By

[†]Equal Contribution

using LLMs to detect intertextuality in biblical writings, the user can establish patterns of usage by individual biblical authors and shed light on how texts were understood and reinterpreted by later writers and illuminate the understanding transmitted by the author from inherited concepts.

Armed with an appropriate pattern of LLM usage, researchers may be able to scale their efforts and shed light on long-standing questions in biblical studies. Further, the methodology can be applied to texts and contexts outside the biblical corpus.

In the following sections we provide the necessary background regarding LLMs and intertextual analysis, discuss the experiments used to substantiate the contributions in Table 1, and describe the experimental results. We conclude with a discussion of the limitations of this work.

2 Background

Richard Hays' seminal work, *Echoes of Scripture in the Letters of Paul*, established intertextuality as a central concern of biblical interpretation for subsequent research (Hays, 1989). The interest in biblical intertextuality has only grown through the years, yet a lack of standardized definitions and methodology remains (Emadi, 2015).

Discerning intertextual relationships in biblical texts is not always a straightforward endeavor. Biblical texts often reference previous writings, both biblical and extra-biblical. Biblical authors are frequently creative in how they use other texts and do not always cite their sources. Often biblical authors will merely allude to, paraphrase, or creatively embed intertextual elements in their writing. This makes intertextual studies challenging and, at times, controversial.

The study of biblical intertextuality is relevant to understanding the relationship of the Old Testament to the New Testament, how later biblical writers understood and interpreted previous texts, and how much contemporary writers knew of other current writings. One prominent area of intertextual study is the synoptic problem, which addresses the relationships of the three synoptic gospels to one another in relation to their interdependence, sources, and manner of composition. Intertextuality has also paved the way for fresh analyses of the relationship between Paul and Jesus (Wenham, 1995). Advances in intertextual analysis through AI and LLMs have the potential to illuminate such

important issues in biblical studies.

2.1 LLMs

Transformer-based large language models (LLMs) like Claude (Anthropic, 2024) have been shown to be remarkably capable of language analysis. LLMs function by generating context aware embeddings of the tokenized words in the context and generating an attention-filtered compression given the current last word in the sequence. The compression is then fed to a non-linear function in the form of a neural network. This process of projection, attention, compression, and non-linear mapping is repeated serially through the layers of the LLM.

Projection into an embedding space has been shown to permit semantic and syntactic reasoning (Mikolov, 2013) through a continuous vector representation. Further, attention has been shown to enable more long range dependencies and meanings to be captured and conveyed into the compressed representation of input text (Bahdanau, 2014). Finally, the feed-forward non-linear mapping provides more efficient learning from the available data (Vaswani, 2017) and universal computation over the input text (Roberts, 2023).

While the technological innovation associated with transformers (Vaswani, 2017) is arguably an increment over natural language processing (NLP) techniques that introduced word vectors (Mikolov, 2013) and attention (Bahdanau, 2014), LLMs have proved to be a revolutionary advance in the performance of virtually all NLP tasks.

2.1.1 Related Work

In Coffee et al. (2012), the authors specifically sought to detect textual allusions by identifying shared words between two texts. They found this method was able to identify previously uncatalogued passages that may contain allusions. While this work was an important step, it is brittle to superficial word overlap and does not have the ability to reason about more nuanced contact between the texts (themes, synonyms, etc).

Dai et al. (2023) and Yu et al. (2024) used GPT-3.5 and GPT-4 (Radford et al., 2019) to do thematic coding. In each, the authors compare the theme label generated by the LLM to the label generated by a human annotators. They show that the tested LLMs tend to be able to reason over thematic content. Finally, Khan et al. (2024) performed LLM assisted corpus coding for function-to-form pragmatic and discourse analysis.

The existing work demonstrates compelling progress toward scalable, automated reasoning for qualitative textual analysis with human collaboration. However, our work importantly augments the existing literature in four ways: 1) direct qualitative intertextual comparison by an LLM, 2) the task requires retrieval rather than labeling, 3) the prompt pattern is intended to augment rather than automate human ability, and 4) texts are presented in a non-English language (Koine Greek).

3 Definitions and Criteria

This paper is concerned with identifying and retrieving specific literary dependencies between biblical texts in Koine Greek using Claude Opus (Anthropic, 2024) which can be evaluated by the expert-in-the-loop to understand intertextual dependence and weight. Due to the multivalent nature of biblical intertextuality, the dependencies may appear as direct quotations, allusions, or echoes (Hays, 1989). Allusion can be defined as an indirect reference with some definable characteristics, such as lexical or thematic correspondence. Echoes, however, are subtle references that may exist purely on a structural or conceptual level or a single specific keyword. While direct quotations are usually apparent to the reader, allusions and echoes require some criteria to help with identification.

Hays (1989) enumerates seven criteria for determining allusions and echoes: availability of the source to the author, volume (characterized by repetition, distinct patterns, prominence), recurrence of the citation by the author, thematic coherence, historical plausibility, history of interpretation, and satisfaction or sensibility. For this study, we will use Hays’ criteria as the expert lens for human evaluation of candidate intertextualities identified by the LLM.

4 Methodology

We identify 6 passages each representing a unique scenario of interest. We provide them to Claude Opus via the pattern in Figure 2 and receive responses similar to that shown. The returned candidate intertextualities are then evaluated for strength based on the criteria discussed above.

In support of the contributions listed in Table 1: We 1) evaluate the effect of query and corpus length by varying each while maintaining the passage and prompt pattern. We 2) evaluate all passages in Koine Greek with an author specialized to intertext-

tual biblical analysis as the expert-in-the-loop providing evaluation over the candidates in the form of a fuzzy weight (labeled as low, mid, or high) as alluded to in Figure 1. We 3,4, and 5) identify failure and performance modes across a comprehensive set of scenarios that present passages with expected false positive, known true positive, probable, speculative, hypothetical, and out-of-sample true positive intertextualities.

Expert

I am looking for intertextualities, that is quotations, allusions, or general dependence on another passage, between two passages. Identify rare words that are mutually used as well. Please denote potential intertextualities between the corpus and query passages below:
[Corpus]
[Query]

LLM

Here are the potential intertextualities I identified between the two passages:

1. The phrase “ἐν ᾧ μέτρῳ μετρεῖτε μετρηθήσεται ὑμῖν” in the corpus passage is very similar to “ᾧ μέτρῳ μετρεῖτε, ἐν αὐτῷ μετρηθήσεται ὑμῖν” in the query. This appears to be a direct quotation.
2. Both passages use the relatively uncommon future passive verbs “κριθήσεσθε” (you will be judged) and “μετρηθήσεται” (it will be measured). ...
5. Both exhort the audience to act/treat others in a certain way, using the 2nd person plural present active verbs ποιεῖτε (you do/make) and δίδοτε (you give).

Figure 2: Prompt pattern used to identify intertextualities between the corpus (to be searched) and the query (passage of interest).

Table 2: Test Scenarios and corresponding corpus and query passages.

Test Scenario	Corpus	Query
True Positive	Matthew 7	1 Clement 13
False Positive	Matthew 7	1 Peter 2:4-8
Probable	Sirach 51	Matthew 11:25-30
Speculative	Romans 3:19-5:11	Luke 18:9-14
Hypothesis 1	Matthew 10:5-42	Acts 20:17-35
Hypothesis 2	Matthew 10:5-42	Luke 10:2-12
Out-of-Sample	Matthew 7	Fake Biblical Stylized Text

4.1 Selected Texts and Scenarios

Matthew 7 is tested against two texts, a known positive and a known passage with superficial similarity, to evaluate the model’s sensitivity in the cases of true positive and expected false positive. In 1 Clement 13, the author directly quotes from Matthew 7:1-2 and attributes the quote to Jesus. 1 Peter 2:4-8 contains no known parallel with Matthew 7 but shares a common stone metaphor with Matthew 7:24-27.

The remaining four sets are all experimental or hypothetical in some way. First, similar to the true positive, some scholars argue that Matthew 11:25-30 depends on Sirach 51 while others merely acknowledge the similarities (Hagner, 1993). On the other hand, most scholars reject dependence between Romans and Luke 18:9-14 but acknowledge superficial thematic similarity (Johnson, 1991).

We suspect Matthew 10:5-42 to have literary interdependence with Acts 20:17-35 based on novel, unpublished research. However, no known proponent of this relationship exists. Interestingly, the implications of this scenario can potentially inform the debate surrounding the synoptic problem.

Finally, to test if LLMs can identify truly novel intertextualities apart from potential knowledge from an unknown pretraining corpus, we develop a novel passage that shares features with Matthew 7 regarding a fish and a tree.

While intertextuality can exist across texts composed in different languages, we chose texts composed in Koine Greek to keep the project manageable. Since intertextuality often depends on lexical forms, unique vocabulary, and morphological features, we conducted the exercises in Greek rather than English translations. Further, research in the field of new testament studies is conducted in Greek, and we hope LLMs will be used to augment the work of other researchers in this space.

5 Experiments

5.1 True Positive: Matthew 7 & 1 Clement 13

Background: In 1 Clement 13, the author seems to synthesize several related sayings that can be found throughout the Sermon on the Mount (Holmes, 2007). Scholars date the writing of 1 Clement as being after or contemporary with the composition of Matthew’s gospel. While it is possible that the authors of both texts draw from a hypothetical common source, no such document exists, and the grouping of the same sayings in both texts suggests some direct dependence (Holmes, 2007).

Candidate Intertextualities: The language model identified the phrase, “With the measure you use, it will be measured to you,” as a direct quote (Matt. 7:2). In Greek, the only variation from the Matthean text is the omission of the preposition ἐν (“by the measure you use”) and the addition of the prepositional phrase ἐν αὐτῷ (“by this it will be measured”). The model detected other lexical parallels, such as the use of the verb κρίνετε “you judge” and κρινθήσεσθε “you will be judged,” as well as the usage of δίδωμι “give” (Matt. 7:7).

The model also detected morphological similarities in verbs used, such as the use of the future passive forms μετρηθήσεται “it will be measured” and κρινθήσεσθε “you will be judged.” It also detected identical imperative forms in person, number, tense, and voice: ποιεῖτε (you do) and δίδοτε (you give). The model detected general thematic and structural features common to both texts.

Observations & Analysis: The 1 Clement citation explicitly credits Jesus with the sayings (“Let us remember the words of the Lord Jesus”). Interestingly, the model mentioned sayings not included in Matthew chapter 7 but did not locate them. This indicates that the model was working within the boundaries of the provided texts but utilizing broader knowledge acquired during pretraining. When we reran the exercise and omitted the explicit reference to Jesus, the model did not acknowledge the additional sayings. The language model generated no false positives and the weight of the intertextual dependence is high.

5.2 False Positive: Matthew 7 & 1 Peter 2:4-8

Background: 1 Peter bears no evidence of direct literary dependence on the synoptic gospels. However, both Matthew 7:24-27 and 1 Peter 2:4-8 use stone metaphors. This provides an opportunity to identify model sensitivity to false positives. As ex-

pected, the model retrieved the superficial textual similarities.

Candidate Intertextualities: The model concluded that 1 Peter 2:4-8 drew on the imagery and language of Matthew 7. It cited the use of construction metaphors, the use of πέτρα “rock/stone” and προσκόπτω and πρόσκομμα “stumble/stumbling stone,” and the theme of “two ways” in both passages. However, the model’s assertion of direct dependence can be confirmed as a false positive since the pertinent references in 1 Peter 2:4-8 are explicit quotations from the Old Testament (Marcar, 2016).

Observations & Analysis: While the model provided a false positive, the exercise was helpful. First, the model built a convincing case with several concrete and sound data points. There are striking similarities between Matthew 7 and 1 Peter 2:4-8. Additionally, the model identified the motif of a rejected stone in both passages. This is the explicit content of the Psalm 118 citation in 1 Peter 2:7, but the model detected the same motif in Matthew 7:9.

However, the shared Old Testament tradition best accounts for these similarities since the stone motif from Isaiah and Psalms is extensively cited in the New Testament. This gives warrant to further research into the potential intertextual link between the Isaiah and Psalms texts with Matthew 7. The language model generated valid candidates with no false positives. The weight of the intertextual dependence is low, but likelihood of shared dependence with a common pretext is mid to high.

5.3 Probable: Sirach 51 & Matthew 11:25-30

Background: The Wisdom of Ben-Sirach, or Sirach, is a second-century B.C.E. Jewish wisdom book in the Deutero-canonical (or apocryphal) writings. Some scholars maintain that Sirach 51 provides the background to Jesus’ words in Matthew 11:25-30, but it is not a universally affirmed intertextual connection (Hagner, 1993). This experiment will confirm whether LLMs can contribute to a positive case for intertextuality between these writings.

Candidate Intertextualities: The model listed several anticipated intertextual links, such as the theme of revealed wisdom, references to the “yoke” of wisdom, instruction, and finding rest. These were valid intertextual candidates and are supported by prominent commentators (Luz and Crouch, 2001). The most meaningful contribution was related to the structure of the passages. The model

noted that both passages begin with thanksgiving and an acknowledgment of God as father (Sirach 51:1, 10; Matthew 11:25-27). It further elaborated that the rhetorical form consisting of a prayer to God followed by an exhortation to seek wisdom was an established pattern in Jewish wisdom literature. The LLM neither confirmed nor denied a direct intertextual relationship between the corpus and query but suggested that the passages draw on similar rhetorical patterns.

Observations & Analysis: Our interest turned to the rhetorical form of the passages and where else this form might occur. When prompted, the LLM provided examples of biblical and extra-biblical texts that reflect this pattern and cited relevant existing research. In this way, the model generated candidate texts for further study. We are unaware of anyone making this distinction regarding the relationship between Sirach 51 and Matthew 11:25-30. Also, there appears to be no established nomenclature for this rhetorical phenomenon.

Further research could shed light not only on the question of intertextuality but also on the literary unity of Sirach 51. This chapter is composed of three poems, and the question of its unity and history of composition is unsettled (Goodman et al., 2012). The insights of the language model have revealed a connection between issues of intertextuality, Jewish wisdom rhetoric, and literary composition, which merits further research.

The language model generated valid candidates with no false positives and the weight of the intertextual dependence is mid.

5.4 Speculative: Romans 3:19-5:11 & Luke 18:9-14

Background: This experiment involves the relationship of Jesus’ sayings to the teachings of Paul. Intertextual analysis of Jesus’ sayings and Paul’s writings has produced fruitful results, leading Wenham (1995), Allison (1982), and others to posit a close connection between Jesus and Paul. The parable in Luke 18:9-14 tells us that the sinful tax collector was “justified” while the Pharisee was not. Here, Luke uses the verb δικαίωω “justify” in a similar manner as Paul when he writes of justification by faith in Romans. The scholarly consensus is that while there is an overlap of ideas, there is no literary interdependence between these two texts (Johnson, 1991). We chose these texts to learn whether LLMs could provide evidence of intertextuality.

The corpus passage, Romans 3:19-5:11, was cho-

Acts 20:17-35	Connection type	Matthew 10:5-42	Connection type	Luke 10:2-12
19, 23 Paul's recounts his suffering	Thematic	17-19, 23 Jesus promises suffering		
22-23 Holy Spirit/persecution	Lexical & Thematic	20 Holy Spirit/persecution		
23 in every city	Lexical & Thematic	5, 11, 14-15 whatever city you enter	Lexical & Thematic	8, 10, 12 whenever you enter a city
24 Paul counts his life of no value	Lexical & Thematic	39 Whoever loses his life finds it		
24 testify	Lexical	18 testify		
25 proclaiming the kingdom	Lexical	7 proclaim... the kingdom...	Lexical	9 the kingdom has come
28-29 church as flock	Thematic	6 Israel as sheep		
29 wolves among sheep	Lexical & Thematic	16 sheep among wolves	Lexical & Thematic	3 sheep among wolves
33 silver, gold, apparel	Lexical & Thematic	9-10 gold, silver... two tunics	Thematic	4 moneybag, knapsack, sandals
35 Paul's hard work	Thematic	10 worker worthy of food	Thematic	7 worker worthy of wages
35 help the weak ἀσθενούντων	Lexical	8 heal the sick ἀσθενούντας	Lexical	9 heal the sick ἀσθενούντας
35 more blessed to give than to receive	Lexical	8 freely you have received, freely give		

Table 3: Candidate Intertextualities for Acts 20:17-35 to Matthew 10:5-42 and Luke 10:2-12 to Matthew 10:5-42

sen based on the distribution of the word δικαιόω in Romans. The word occurs 15 times in Romans; the selected portion contains nine occurrences. When we asked the model to narrow the corpus passage to the most concentrated section of potential intertextual connections, it identified Romans 3:21-4:8 as the portion with the highest density and differentiated it from the subsequent paragraph about Abraham. Seven of the nine occurrences of δικαιόω in the original corpus occur within the narrowed corpus. By narrowing the corpus, the LLM conducted what Guthrie (1993) calls "cohesion shift analysis" which detects shifts in "cohesion fields" usually around paragraph breaks.

Candidate Intertextualities: The LLM picked up on the repeated key terms δικαιόω "justify" and ἁμαρτωλός "sinner" as well as their cognate noun forms δικαιοσύνη "righteousness/justice" and ἁμαρτία "sin" which occur frequently in the corpus. The LLM made loose thematic connections as well, such as the central role faith plays in both passages (although the Greek word πίστις is not used in Luke 18:9-14). The model rightly noted that both passages feature righteousness apart from works or boasting. The Pharisee in Luke 18:9-14 illustrates the concept which is explicit in Romans 3:21-4:8.

The LLM rightly notes the occurrence of a rare word in both passages. In Romans 3:25, Christ is a ἱλαστήριον "propitiation" for sin, and in Luke 18:13, the tax collector asks God to ἰλάσθητί "propitiate" him, the sinner (ἁμαρτωλός). The words are cognates: a noun in Romans and a verb in Luke. Both words only occur twice in the New Testament

(the other occurrences are in the Epistle to the Hebrews). The LLM curiously cited references to Abraham as an intertextual candidate, while there is no reference to Abraham in Luke 18:9-14.

We then asked the LLM to locate potential intertextual references to Romans 3:21-4:8 in the preceding and subsequent pericopes. Using Luke 18:1-8 as the new query, the LLM found two lexical correspondences. The use of the word πίστις "faith" in Luke 18:8 and two cognate words with δικαιόω/δικαιοσύνη, ἐκδίκησόν meaning "grant justice" and ἀδικίας which describes the "unrighteous" judge.

Next, we asked the LLM to analyze Luke 18:15-27, which follows the original query passage. It provided three meaningful connections. It juxtaposed the emphasis on righteousness apart from the law in Romans to the focus on keeping the law in Luke 18:20-21. The LLM also noted the correspondence between the righteousness of God in Romans and the statement in Luke 18:19 that "none is good except God alone." The LLM correlated the statement in Luke 18:27 that nothing is impossible with God with the claim in Romans 4:5 that God justifies the ungodly, which is impossible with man.

Observations & Analysis:

While this exercise did not yield groundbreaking insights, the LLM provided valuable analysis of the passages in question and handled multiple queries well. When asked to evaluate the three queries in terms of their relatedness to the corpus, the model identified the original query as the most

related. However, at no point did the LLM assert direct literary dependence between the corpus and queries.

The language model generated valid candidates with one false positive and the weight of the intertextual dependence is low to mid.

5.5 Hypothesis: Matthew 10:5-42, Acts 20:17-35, & Luke 10:2-12

Background: The rationale behind this exercise is exploratory. While working on another project, we found a journal article that noted similarities between Paul's farewell discourse in Acts 20:17-35 and Jesus' commissioning of the disciples (Brown, 1963). Upon analyzing the passages, we suspected direct literary dependence between Matthew 10:5-42 and Acts 20:17-35 though no known scholarship explores intertextual connections between these passages. Since the passages in Matthew and Luke are parallel passages, we ran both against Acts 20:17-35 to see which had the strongest connections. Interestingly, Acts 20:17-35 and Luke 10:2-12 do not appear to have as strong of an intertextual dependence even though they share an author.

Observations & Analysis: While the model acknowledged the possibility of literary dependence with both, it cited six intertextualities with Luke and twelve with Matthew. This is mainly due to Matthew's expanded version of the missionary discourse. Matthew's version contains unique material not included in Luke but alluded to in the Acts discourse. This is a remarkable observation considering that Luke and Acts have the same author, and Luke's version of the missionary discourse omits much of the Matthean material present in Acts 20:17-35.

When using Acts 20:17-35 as the query passage, the LLM provided several parallels and concluded that the passage was dependent on the Matthew passage. However, we continued submitting prompts and encountered false information and non-sequitur reasoning. We determined that this was due to the query size. We broke the query passage into smaller sections and ran each section separately. Reducing the query size also reduced the number of parallels, and the LLM did not recognize patterns of literary dependence when working through the query one section at a time.

The language model generated over a dozen valid candidates with some false positives and the weight of the intertextual dependence is high.

5.6 Out-of-Sample: Matthew 7 & Fish and the Tree

Background: This comparison is intended to evaluate the LLM's ability to retrieve intertextualities when it is guaranteed that the model has no background information from pre-training on which to draw intertextual information. A novel, moralistic story stylized via an LLM was written in Koine Greek involving a tree and a fish. This story intentionally included motifs and words from the Matthean passage.

Candidate Intertextualities: The model generated the following intertextual candidates: the metaphor of a tree bearing fruit, the word ἰχθύς (fish), the thematic similarity between deep roots and a rock foundation, and the thematic similarity between deep roots of faith and good fruits of faith. The model generated two false positives by identifying the words ποταμός (river) and ἀγαθός (good) as rare words. The word ἀγαθός occurs 101 times in the New Testament while ποταμός occurs 17 times.

Observations & Analysis: The model correctly identified intertextual similarities even though the query passage was entirely out of context. This shows with certainty that LLMs are capable of intertextual candidate retrieval without apriori knowledge.

English Translation of the Parable of the Fish & the Tree: The river flows through the forest, and by the river there is a great tree. But the tree bore good fruit and gave shade to the animals of the forest. But in the river there is a small fish, which asks for food every day. And the fish saw the tree and its fruits falling into the water. And he ate of the fruit and gave thanks to the tree. But there was a great flood and the river was flooded. But the river dragged the fish away from the tree. And the fish was troubled and afraid, having no food or shelter. But the tree had deep and strong roots, and it remained firm in the flood. And when the water receded, the fish found its way to the tree again. And they lived in peace, the tree providing and the fish giving thanks. Likewise we, if we have a deep root of faith, remain firm in tribulations and will find the way to God again. Because He always provides us with His goods according to the measure of our faith.

6 Results and Observations

The LLM successfully identified lexical correspondences by detecting common words and analyzing morphological data in verbal forms, including tense, mood, voice, person, and number. It also provided keyword statistics, sometimes identifying rare words and the number of occurrences in the New Testament. The model identified direct quotes, even when the quoted form was adapted or paraphrased. It made an important distinction between direct quotes and verbatim citations. Discerning a direct quote not in verbatim form shows advanced language processing capability. In addition to detecting these textual phenomena, the LLM also identified areas where these features were most dense within a larger corpus.

In addition to lexical and textual analysis, the LLM found intertextual relationships through other modes of analysis. The model performed contextual analysis of the passages. It could differentiate between the usage of the same word in two different contexts. The same word used in a different context from the corpus text was not afforded the same weight as a word used in a similar or identical context. Additionally, the model could detect thematic correspondence when no lexical parallel was present. It also detected shared structural features that indicate relatedness between texts.

It is worth noting that the level of analysis conducted by the LLM has great potential beyond intertextual study. The type of structural, lexical, and morphological analysis modeled here has been productively used for discourse analysis (Umphrey, 2022). Like intertextual studies, discourse analysis suffers from a lack of clear methodological consensus and could benefit from the advancements offered by LLMs.

6.1 Evidence of Novel Intertextual Work

One concern of this type of experiment is that the language model would draw from background knowledge in the training data rather than conducting novel analyses of the provided texts. On multiple occasions, the LLM demonstrated the ability to reason exclusively from provided texts yet draw from its knowledge bank as necessary. When analyzing the sayings of Jesus in 1 Clement 13, it acknowledged that some of the sayings were from outside the corpus text of Matthew 7 but did recognize the previous chapters of the Sermon on the Mount as their source (Matthew 5-7). This indi-

cates that the LLM was doing a closed analysis of the provided texts. Similarly, the LLM asserted intertextual dependence between 1 Peter 2:4-8 and Matthew 7, when the correspondence points are direct quotations from Isaiah and Psalms. The model reasoned from the provided texts without looking outside them. This suggests the LLM tended to limit the scope of consideration, avoiding irrelevant retrieval beyond the target passage.

The analysis of Sirach 51 provided what appears to be a completely novel intertextual observation regarding the structure of both passages. Furthermore, the intertextual links discovered in the analysis of Acts and Matthew seem to be without precedent in the scholarly literature. These observations have potential for future research. We are convinced that LLMs have great potential for generating novel research ideas for biblical intertextual studies.

Finally, by retrieving intertextual candidates for an unseen passage, the LLM has shown that it is able to reason over complex intertextualities without the benefit of pre-trained knowledge.

7 Conclusions

This study demonstrates the potential for using LLMs to identify and examine intertextual relationships within biblical texts. By evaluating the effect of query and corpus length, testing performance on passages in Koine Greek, and assessing the model's ability to handle various intertextuality scenarios, we have shown that LLMs can be a valuable tool for biblical scholars. This tool works exceptionally well when the query is short and the corpus is between 1 and 3 chapters. By successive application of the pattern across corpora, this method can be used to evaluate large bodies of texts for intertextual connections to a query passage in a scalable manner.

Our findings suggest that LLMs are capable of detecting direct quotations, allusions, and echoes between biblical texts, even when presented in a non-English language. The LLM successfully identified lexical correspondences, morphological similarities, direct quotes (even when adapted or paraphrased), and thematic and structural parallels. It also demonstrated the ability to narrow down a corpus to the most relevant sections for intertextual analysis.

Importantly, the LLM exhibited evidence of novel intertextual work, generating observations

and connections that appear to be unprecedented in the scholarly literature. This suggests that LLMs have the potential to uncover new insights and generate fresh ideas for biblical intertextual studies.

7.1 Future Work

An interesting facet of this paper is the inspiration from fuzzy neural networks which provided an important lens. However, an important limitation of any work that attempts to characterize the intertextual dependence weight between any pair of passages is that it will necessarily fail to determine relational direction since historical information is sparse. So, future work should simultaneously consider multiple nodes within the graph to identify the order of intertextual dependence which finds the most probable candidate textual chain.

Future work should also investigate the impact of contextual recall effects like the fan effect observed in some LLMs (Roberts et al., 2024). It may be that objects or people which appear frequently in the corpus in varying scenarios may be more highly impacted by LLM hallucinations leading to a greater frequency of intertextual false positives.

8 Limitations

There were some limitations in the LLM's intertextual analysis which must be considered. The model struggled with long query passages, occasionally producing errors and non-sequiturs. It also failed to consider shared pretexts in some cases, asserting direct dependence between texts when a common source was more likely. Additionally, the intertextual candidates generated by the LLM sometimes included false positives or tenuous connections, requiring expert evaluation.

First, the model does not perform well with long queries. In the experiment with Acts 20 and Matthew 10, the model produced basic errors such as identifying words that did not exist and making non-sequitur judgments. The errors compounded as more prompts were submitted.

Next, when presented with two similar texts for analysis, the model may not consider a shared pretext, even if one exists. This was the case with the analysis of Matthew 7 and 1 Peter 2:4-8. The model asserted direct dependence between the query and corpus without considering the shared pretexts in Psalms and Isaiah.

Finally, the candidates generated by AI require the critical eye of an expert in the field. Of Hays'

seven criteria for intertextuality used for this paper, the LLM provided results based on volume, thematic coherence, and sensibility. The LLM did not evaluate availability, recurrence, historical plausibility, or the history of interpretation. The user must possess these competencies to properly evaluate generated candidates.

References

- Graham Allen. 2011. *Intertextuality*. routledge.
- Dale C Allison. 1982. The pauline epistles and the synoptic gospels: The pattern of the parallels. *New Testament Studies*, 28(1):1–32.
- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.
- Dzmitry Bahdanau. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- John Pairman Brown. 1963. Synoptic parallels in the epistles and form-history. *New Testament Studies*, 10(1):27–48.
- Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W Forstall, Roelant Ossewaarde, and Sarah L Jacobson. 2012. The tesserae project: intertextual analysis of latin poetry. *Literary and linguistic computing*, 28(2):221–228.
- Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. [LLM-in-the-loop: Leveraging large language model for thematic analysis](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore. Association for Computational Linguistics.
- Samuel Emadi. 2015. Intertextuality in new testament scholarship: Significance, criteria, and the art of intertextual reading. *Currents in Biblical Research*, 14(1):8–23.
- Martin Goodman, John Barton, and John Muddiman. 2012. *The Apocrypha*. OUP Oxford.
- George H Guthrie. 1993. *The structure of Hebrews: A text-linguistic analysis*, volume 73. Brill.
- Donald A. Hagner. 1993. *Matthew 1–13*, volume 33A of *Word Biblical Commentary*. Word, Incorporated, Dallas.
- Richard B Hays. 1989. *Echoes of Scripture in the Letters of Paul*. Yale University Press.
- Michael W Holmes. 2007. Apostolic fathers: Greek texts and english translations. *Baker Academic*.
- Luke Timothy Johnson. 1991. *The gospel of Luke*. 3. Liturgical Press.

- Awais Hameed Khan, Hiruni Kegalle, Rhea D'Silva, Ned Watt, Daniel Whelan-Shamy, Lida Ghahremanlou, and Liam Magee. 2024. Automating thematic analysis: How llms analyse controversial topics. *arXiv preprint arXiv:2405.06919*.
- Julia Kristeva. 1980. Desire in language: a semiotic approach to literature and art. *Columbia UP*.
- Iliia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.
- Hon Keung Kwan and Yaling Cai. 1994. A fuzzy neural network and its application to pattern recognition. *IEEE transactions on Fuzzy Systems*, 2(3):185–193.
- Ulrich Luz and James Crouch. 2001. *Matthew 8-20: Hermeneia series*. Fortress Press.
- Katie Marcar. 2016. The quotations of isaiah in 1 peter: A text-critical analysis. *TC: A Journal of Biblical Textual Criticism*, 21:1–22.
- Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- John Milton. 2005. *Paradise lost*. Hackett Publishing.
- Kenny Ortega. 1993. *Hocus pocus*. Motion picture. United States.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Jesse Roberts. 2023. How powerful are decoder-only transformer neural models? *arXiv preprint arXiv:2305.17026*.
- Jesse Roberts, Kyle Moore, Thao Pham, Oseremhen Ewaleifoh, and Doug Fisher. 2024. Large language model recall uncertainty is modulated by the fan effect. *arXiv preprint arXiv:2407.06349*.
- Ray Arlin Umphrey. 2022. *The Cohesive Rhetoric of Jesus: The Role of Matthew 6: 19–7: 11 in the Sermon on the Mount*. The Southern Baptist Theological Seminary.
- A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- David Wenham. 1995. *Paul: follower of Jesus or founder of Christianity?* Eerdmans.
- Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*.

Extracting Relations from Ecclesiastical Cultural Heritage Texts

Giulia Cruciani

Università degli Studi di Messina / Messina, Italy
giulia.cruciani@studenti.unime.it

Abstract

Motivated by the increasing volume of data and the necessity of getting valuable insights, this research describes the process of extracting entities and relations from Italian texts in the context of ecclesiastical cultural heritage data. Named Entity Recognition (NER) and Relation Extraction (RE) are paramount tasks in Natural Language Processing. This paper presents a traditional methodology based on a two-step procedure: firstly, a custom model for Named Entity Recognition extracts entities from data, and then, a multi-input neural network model is trained to perform Relation Classification as a multi-label classification problem. Data are provided by IDS&Unitelm (technological partner of the IT Services and National Office for Ecclesiastical Cultural Heritage and Religious Buildings of CEI, the Italian Episcopal Conference) and concerns biographical texts of 9,982 entities of type person, which can be accessed by the online portal BeWeb. This approach aims to enhance the organization and accessibility of ecclesiastical cultural heritage data, offering deeper insights into historical biographical records.

1 Introduction

In the current landscape, the abundance of data is unprecedented; technological advancements, the Internet of Things (IoT), increasing connectivity, and digitalization are some factors that led to today's scenario. Yet, whereas methods of data collection continue to expand, the true value of this phenomenon is not related to the mere accumulation of information, but to the acquisition of meaningful insights. For this reason, there is a major focus on applying innovative techniques to all possible domains; the cultural heritage environment is also experimenting with big attention to exploit new possibilities. In the post-COVID era, the National Recovery and Resilience Plan (NRRP) has funded numerous endeavors to underline the impor-

tance of preserving and exploring cultural heritage, signaling a great moment for the combination of innovative approaches with domains traditionally less associated with these topics. This study positions itself among those initiatives, aiming to bridge this gap by employing advanced methodologies within the cultural heritage landscape. At the core of this pursuit, a paramount role is played by Knowledge Graphs, tools based on the concept of Knowledge Bases. Knowledge Graphs enable us to navigate the complex and intricate depths of data; for this reason, this work aims at finding the basic components of a knowledge graph, by extracting entities and relations from texts. The proposed methodology uses a traditional approach structured as a two-step process: first Named Entity Recognition (NER), key for pinpointing names, locations, and other text elements, and later sentence-level Relation Extraction, implemented as a multi-class classification task. A multi-input neural network model is built to leverage a labeled dataset of sentences and entity types. In Natural Language Processing, extracting semantic relationships from text is a very crucial task. This process concerns converting unstructured data (text) into structured. Relation Extraction (RE) can be achieved in several ways; one of them consists of setting the problem as a Classification task: Relation Classification (RC) (Zeng et al., 2014), (Zhang et al., 2017). Relation Classification has been approached using pattern-based (Suchanek et al., 2006), (Kambhatla, 2004), or kernel-based methods (Zhou et al., 2016). Early approaches make use of pipelines that identify entities and then classify relations between pairs using CNNs or LSTMs to capture sentence-level semantics (Zeng et al., 2014), (Zhou et al., 2016). Moreover, current methods for sentence-level Relation Extraction employ Transformer models (Yamada et al., 2020) like BERT (Devlin et al., 2019). In the context of art, especially in the last years, increasing importance is reserved for applications of digital tools

mainly focused on art discovery or recommendation mechanisms (Gonzalez and Andrew, 2014). In Santini et al. (2022) the authors describe several techniques (such as entity recognition and linking, coreference resolution, time extraction, and artwork extraction) applied to Vasari’s most important piece “The Lives of the Artists”. In Chen et al. (2022), instead, NLP techniques have been applied to biographical texts of artists to achieve sentence-pair binary classification, hence connections among artists, without considering the type of the relations, conversely to what is developed in this study. The paper is organized as follows: sections 2 and 3 respectively illustrate the data and the methodology that were employed, starting from the formulation of the problem and continuing with the construction of models, and their evaluation (section 4). The last part shows conclusions and potential future applications.

2 Data

Data for this research were provided by the partner company ¹ and can be accessed through an online website: BeWeb (<https://beweb.chiesacattolica.it/>). The portal allows seamless exploration of diverse databases acquired through the joint efforts of UNBCE ² and the Italian dioceses (Russo, 2014). This collaborative initiative began in 1996 and has resulted in a census of diocesan and ecclesiastical assets, including archives, libraries, and museums. Throughout the years they were able to assemble a database that comprises over 5 million records, including 4 million historical and artistic assets, 66,000 places of worship, 1.5 million library assets, 6,800 archival collections, and 1,588 cultural institutions (Weston et al., 2017). Thanks to the help of approximately 3,000 experts, the projects are constantly reviewed and updated. Upon validation, the data is integrated into the national database; BeWeb, therefore, shows data validated from several inventories and facilitates cross-domain navigation of databases making use of the specific descriptive standards for each sector. The project’s development involves two main elements: a dataset derived from sector-

¹This project is the result of a PhD program financed by an EU scholarship initiative designed for Italian public school students, aimed at fostering innovation, and is a collaboration between University of Messina (Italy) and the local company IDS&Unitelm.

²UNBCE stands for Ufficio Nazionale Beni Culturali Ecclesiastici, namely National Office of Ecclesiastical Cultural Assets

specific descriptive standards and a clustering system that interconnects terms referring to the same entity across various catalog databases, resulting in clusters identified by cross-domain aggregates: CEI ³ Authority File. Hence, the CEI Authority File (AF-CEI) can be seen as a centralized repository that integrates authority records from diverse cataloging domains through a clustering mechanism. Each domain interacts with the AF-CEI to establish, associate, or revise clusters. These resulting clusters undergo enrichment by designated reviewers, incorporating additional elements such as alternative nomenclature, biographical and historical annotations, images representing the entity, where present, interrelationships with other AF-CEI, and potential references to external web resources (Weston et al., 2017). This research focuses on the biographical notes of 9,982 CEI Authority Files categorized as Person.

3 Methodology

3.1 Problem Definition

The primary goal of this study is to extract structured triplets from Italian biographical texts. The study concerning entities and relations can be formulated as the exploration of triplets, such as $\{e_1, rel, e_2\}$, where e_1 and e_2 are respectively the first and second entity in the sentence, and rel is the relationship existing between the two. The proposed work, therefore, starts with an initial assumption: given that the text analysis involves biographical notes of entities categorized as “Person”, and considering a sentence-level relation classification, the e_1 in the triplet is always the entity whose texts is being analyzed (namely the CEI Authority File), while e_2 changes each time a new entity is extracted from the sentences. For this reason, these triplets consist of the implicit entity that is the subject of the biographical text (referred to as the Authority File entity), the second entity, which is explicitly mentioned in the sentence and extracted using Named Entity Recognition (NER), and the relationship between the two, which is classified based on predefined relationship categories. The final output of this process will eventually look like: $\{Authority_File, relationship, entity\}$.

³CEI stands for "Conferenza Episcopale Italiana", namely the assembly of bishops of the Catholic Church, responsible for coordinating and promoting the Church’s activities and policies in Italy.

Authority File	Text Note
Franco Margari	Painter, graphic designer, video artist. He trained at the Accademia di Belle Arti of Rome. He began his artistic experience in the graphic field in the 1980s and specialized in engraving techniques; since the early 1990s he has also simultaneously dedicated himself to painting: in 1993 he began his exhibition activity in 1993. In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence. There are numerous presences in public and private collections. He lives and works in Florence.

Table 1: Biographical Text Note of Franco Margari

3.2 Data Segmentation and Entity Type

Named Entity Recognition (NER) plays a crucial role in this research; however, its standalone application is insufficient and some adjustments to the data are imperative to facilitate the sentence-level extraction of relationships. As mentioned above, the aim is to identify triplets where the Authority File itself denotes the first component, the relation is intrinsic in the semantics of the sentence and the third unit encompasses every other entity identified by the model. Consider Table 1 as an illustrative example.

The initial column denotes the name of the Authority File, specifically "Franco Margari", while the subsequent column contains the corresponding text data.⁴ Firstly, texts undergo a segmentation into individual sentences. This segmentation is initiated at every punctuation dot occurrence, marking a new sentence's beginning. The NER model is employed to analyze each sentence within the data and as well as identifying the entities, it is asked to retain only sentences containing named entities, as depicted in Table 2.

Sentences such as "Painter, graphic designer,

⁴Texts are originally in Italian, but for the sake of understanding examples will show an English-translated version.

video artist", "He began his artistic experience in the graphic field in the 1980s and specialized in engraving techniques; since the early 1990s he has also simultaneously dedicated himself to painting: in 1993 he began his exhibition activity in 1993", "There are numerous presences in public and private collections" are excluded as the model did not detect any entities within them. Moreover, in cases where a single sentence contains multiple entities, the model generates an equivalent number of triplets. Table 2 shows an example: the second sentence ("In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence") encompasses two separate entities, namely "Fiorino d'Argento" and "Palazzo Vecchio in Florence". The model extracts both entities, resulting in the creation of not only three triplets - as suggested by the number of sentences - but rather four, accounting for the total number of named entities identified and extracted by the model. For this reason, the total number of potential triplets, hence the total number of relations, is obtained after applying the NER model to the dataset. In addition, while extracting the entities, the model is asked to create another field to account for the type of each entity. This augmentation is motivated by the fact that sentences may include multiple entities, some of which may correspond to distinct types of relations. Notably, the differentiation between entities often underlines the nature of these relations. Hence the augmentation with entity types can offer enhanced insights into the diverse nature of relations, facilitating a more comprehensive understanding of the associations and enabling a more granular analysis of the relationships within the data.

3.3 Identification of relations

The dataset used in this study is highly domain-oriented, consequently, the choice of multi-label classification for relation extraction relies on the fact that recognizing the topic of a sentence will likely reflect the nature of the relationship among the entities mentioned within that sentence. For this reason, four distinct categories were identified to represent all the potential relationships in the texts: Work/Study, Birth/Death/Travel, Kinship, and Ecclesiastical Titles. The classes were constructed intentionally to be broad and able to encompass similar relations. Table 3 shows the categories and relative explanations. After identifying the categories, relation classification is achieved with a multi-label classification model, part of supervised

Authority File	Sentence-Level Note	Extracted Named Entity	Entity Type
Franco Margari	He trained at the Accademia di Belle Arti of Rome	Accademia di Belle Arti of Rome	Organization
Franco Margari	In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence	Fiorino d'Argento	Miscellaneous
Franco Margari	In 2019 he was awarded the Fiorino d'Argento for graphics at Palazzo Vecchio in Florence	Palazzo Vecchio in Florence	Organization
Franco Margari	He lives and works in Florence	Florence	Location

Table 2: Named Entity Recognition on Texts

Relationship	Explanation	Examples
Work/Study	Connections expressing work or study relations.	“Franco Margari trained at the Accademia di Belle Arti of Rome.”
Birth/Death/Travel	A relation between a Person and a Location, that represents the place where the Person was born, found dead, or traveled to.	“Cassiano Carpaneto died in 1998 and was buried in the Langasco cemetery.”
Kinship	Relations showing a familiar bond, such as: “is son of”, “is married to”.	“Giovan Battista Del Tasso, son of Marco di Domenico.”
Ecclesiastical Titles	Religious relationship.	“Piero Novati was a priest of the Diocese of Lodi.”

Table 3: Classes for Relationships and Examples

machine learning. This branch of machine learning involves presenting the algorithm with input data along with the corresponding correct output so that the model can learn patterns and relationships between inputs and outputs. For this purpose, a sample of 1,000 sentences, was randomly selected and used for manual annotation. Table 4 shows an example of one annotated sentence.

Fig. 1 displays the distribution of the categories within the annotated dataset and shows a quite balanced division across all classes. As will be discussed later, the model is trained twice, once on the dataset containing 1,000 labeled sentences, and again on an augmented dataset that comprises 2,000 newly annotated sentences. Hence, the augmented dataset (composed of 3,000 labeled sentences) shows a different distribution for the categories, as depicted in Fig. 2, and a significant imbalance among classes. This problem is addressed later during the compiling of the model, when the optimizer, loss function, and custom metrics are defined.

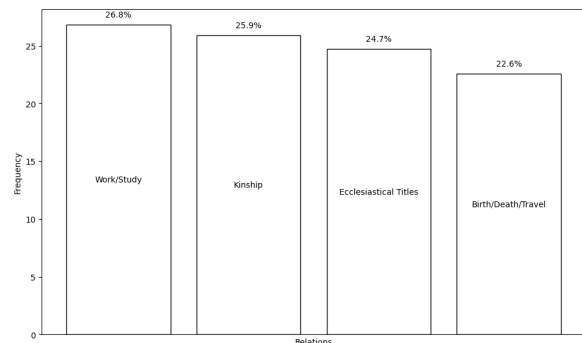


Figure 1: Distribution of relations in the first annotated dataset

3.4 Named Entity Recognition

Named Entity Recognition (NER) techniques aim at identifying significant elements such as names, locations, companies, or events within texts. While various methods exist for NER, this study focused on developing a custom pipeline using the publicly available SpaCy library. SpaCy is an open-source tool that employs machine learning models trained on extensive corpora to detect entities. It offers pre-trained models for different languages and domains. Specifically, the Italian version of

Authority File	Sentence	Extracted Named Entity	Entity Type	Label
Franco Mar-gari	He trained at the Ac-cademia di Belle Arti of Rome	Accademia di Belle Arti of Rome	Organization	Work/Study

Table 4: Annotation Example

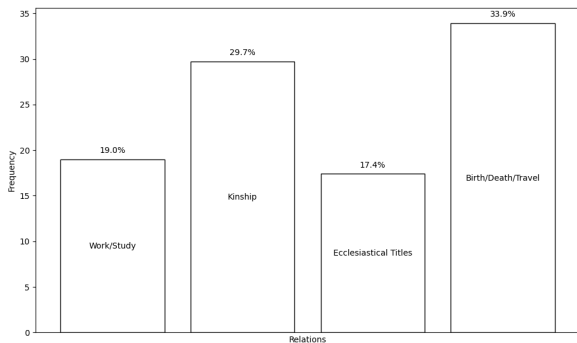


Figure 2: Distribution of relations in the second annotated dataset

SpaCy classifies entities into four categories: person, location, organizations, and miscellaneous; the same notation was adopted in this application, with the only difference being that they were translated into Italian. However, given the domain-specific nature of our dataset, creating a custom model seemed the best choice. Training a custom pipeline using SpaCy is a straightforward process. Detailed guidelines and configuration resources can be found on SpaCy’s training documentation (<https://spacy.io/usage/training>). For the model’s training, a random sample was drawn from the original dataset, consisting of 1,000 biographic annotations, and was split into the standard 70% for training and 30% for testing. Using an on-line platform (<https://tecoholic.github.io/ner-annotator/>), the subset was manually annotated, considering four entity types, mirroring SpaCy’s pre-trained NER model for Italian: Person, Organization, Location, and Miscellaneous. The training phase involved utilizing SpaCy’s provided configuration files; following training, the model’s performance was tested. Fig. 3 depicts evaluation metrics such as loss, precision, recall, and F1-score.

Loss values initially start quite high but gradually decrease over epochs, while all evaluation metrics (F1-score, Precision, and Recall) showcase upward trends over the iterations. The precision and recall values in Fig. 3 reflect only the identification

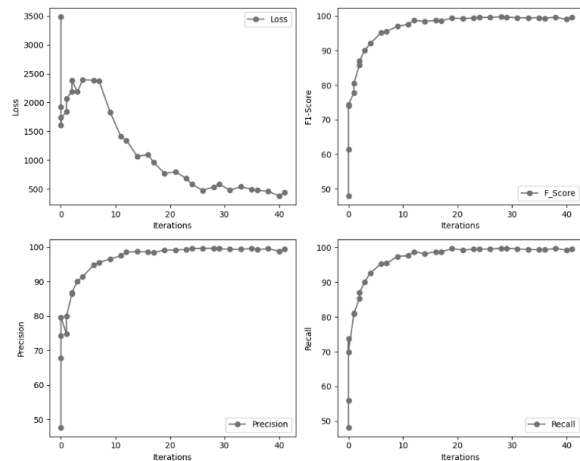


Figure 3: Evaluation Metrics for Named Entity Recognition

of named entities, not their types. These metrics were calculated based on the test set, ensuring the model’s performance was properly evaluated without influencing subsequent training iterations. The number of extracted entities is the following: 4243 Location, 2471 Miscellaneous, 11731 Organization, and 10651 Person. However, it is pertinent to note that these figures are distinct values. A higher number of entities was extracted, but they have been manipulated to account for instances where a singular entity might be referenced multiple times, albeit with slight variations in notation or expression. For this reason, whereas the sum of all distinct entities is 29,098, the total number of triplets resulting from NER is, instead, 65,289.

3.5 Relation Classification

This section introduces a recurrent neural network multi-label classification model constructed using TensorFlow, an open-source framework created by Google specifically for machine learning applications (Abadi et al., 2016); its workflow comprises defining an architecture, compiling, training, and finally evaluating the model. Fig. 4 shows the architecture of the model.

As mentioned previously, the model is constructed by defining two input layers: the sentences

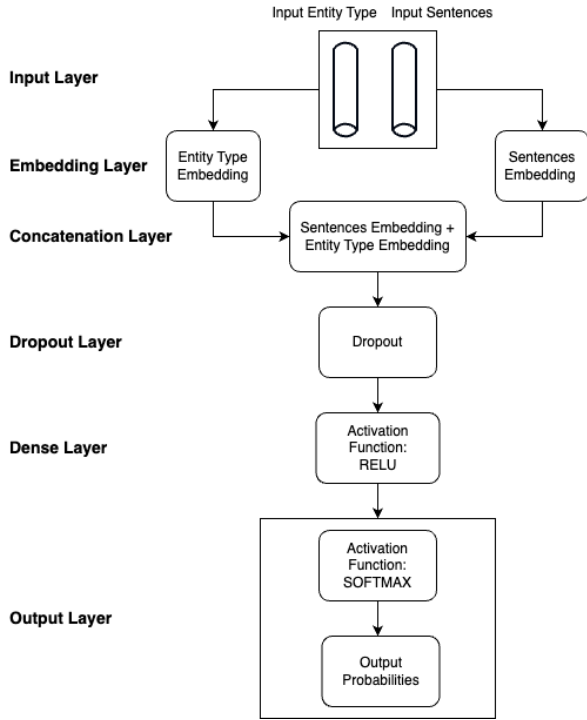


Figure 4: The model’s architecture

and the type of the entities extracted in that specific sentence. The difference in the inputs defines a difference in the approach as well. Sentences, indeed, are passed through the Universal Sentence Encoder (Cer et al., 2018) which preprocesses and then encodes the sentences into embedded vectors. Since entity types, instead, are categorical data they are treated with the so-called one-hot encoding, namely a technique used to transform qualitative features into a format that can be easily comprehended by machine learning algorithms. In this specific case, there are four categories (Person, Location, Organization, Miscellaneous), and applying one-hot encoding results in a binary vector for each. For instance, the class “Person” might be represented as [1, 0, 0, 0] whereas “Location” might be represented as [0, 1, 0, 0] and so on. Each position in the vector corresponds to a category, and only one position has a value of 1, while the others are 0. After obtaining embedding vectors, the two inputs go through a concatenation layer, where they are merged to allow the model to learn from both sources of information simultaneously. This combined representation, then, is fed into a dropout layer, that behaves as a regularization process created to prevent overfitting in neural networks, by dropping a fraction of input units randomly. To that end, during the training a portion of inputs is

randomly set to zero preventing the co-adaptation of neurons; this process encourages the network to learn from robust features that generalize better to unseen data. Given an input x (which in this case is represented by the output from the concatenation of sentence and entity type embeddings) the dropout layer randomly sets a fraction (p) of input units to zero, where p is the dropout rate. In this case, the dropout rate is set to 0.20. Next, a dense layer is introduced, establishing a fully connected hidden layer where the Rectified Linear Unit (ReLU) activation function is applied to the input tensor x . In this layer the input data is transformed, allowing the network to learn and extract higher-level representations and meaningful features from the merged inputs. Mathematically, a ReLU activation function can be defined as in (1); for a certain input x

$$f(x) = \max(0, x) \quad (1)$$

meaning that for any input x , the function will output the maximum of that input’s value or zero. This essentially signifies that the function “activates” (returns a non-zero value) only when the input is positive, and otherwise, remains inactive (returns zero). ReLU is used in neural networks to introduce non-linearity, which enables the network to learn complex patterns and relationships in intricate data. The last dense layer in the model serves as the output layer; its purpose is to produce the final predictions, namely the classifications, based on the features learned from the previous steps. The activation function, in this case, is the SoftMax. Given a vector z of K real numbers (where K is the number of classes, four in this case), the SoftMax function computes the probability $p(y_i)$ for each class i as in (2).

$$p(y_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (2)$$

The numerator in (2) calculates the exponential value of the i th element of z ; this value is then normalized by dividing by the sum of all exponential values. This process forms a probability distribution through the function, that guarantees that the total sum of probabilities across all classes equals 1 and that allows the model to make predictions based on the class with the highest probability; the resulting values represent the likelihood of the input belonging to each class. In the compilation step, the model’s configurations and op-

timizations are determined. For this architecture, the Adam optimizer (Kingma and Ba, 2014) was chosen, along with categorical cross-entropy as the loss function. This combination is commonly regarded as effective for training neural networks, particularly in multi-class classification problems. Adam optimizer is popular due to its effectiveness, adaptability, and efficiency in optimizing the training process for neural networks. The objective during training is to minimize the categorical cross-entropy loss by adjusting the model’s weights and biases using the Adam optimizer. Additionally, specific metrics are defined to monitor both training and validation phases. In this case, the metrics (Precision, Recall, and F1-score) are designed for multi-class scenarios where class imbalance might exist, ensuring that the evaluation considers performance equally across all classes rather than being biased towards majority classes. The model is then trained following two separate phases. Initially, the labeled dataset comprises 1,000 manually annotated sentences, forming the initial training data for the model. The training duration is set to 20 epochs and the model undergoes iterative learning, gradually improving its ability to predict relations between entities within the provided sentences. Upon training the model, it is then applied to a new dataset containing 2,000 sentences for label prediction. Once the model classifies the relations, after a manual analysis where any incorrect predictions are rectified, the corrected ones are appended to the original annotated dataset. This process results in an augmented dataset useful to refine and improve the model. At this point, the model benefits from a larger corpus of labeled sentences and shows a more robust and accurate performance due to the increased diversity and quantity of annotated data. Table 5 illustrates the performance of the two models on the test set, namely a subset that the model does not see during the training phase and that consists of 30% of the annotated datasets.

As depicted in Table 5, the second model outperforms the first significantly in terms of loss, indicating lower errors in predictions, and demonstrates higher accuracy, achieving approximately 89% compared to the first model’s 83%. There are also improvements concerning Balanced Recall, Precision, and F1-score implying enhanced ability in correctly identifying instances across different classes while minimizing false positives. Overall, the second model displays superior performance across all evaluation metrics.

Metrics	First Model	Second Model
Loss	0.47	0.33
Accuracy	0.83	0.89
Balanced Precision	0.85	0.88
Balanced Recall	0.80	0.85
Balanced F1-Score	0.82	0.86

Table 5: Test Sets Evaluation

4 Evaluation of Relationship Classification

To ensure the model grasps relationships effectively without merely memorizing the training data, the evaluation is carried out in two distinct ways. First, evaluation metrics are observed for both training and validation sets during the training process. The validation set is created using 30% of the training data and is used to assess how well the model generalizes to unseen examples. This step aids in identifying any signs of overfitting, where the model might excessively adapt to the training data and struggle to perform well on new data. By validating the model’s performance on this subset, we ensure it can make accurate predictions beyond the examples it was trained on. Initially, the model starts with a loss of 0.9039 and an accuracy of approximately 61.8%, however, as training proceeds, the model shows notable improvements, with decreasing loss values and increasing accuracy, as well as balanced metrics such as Recall, Precision, and F1-score. These metrics exhibit positive trajectories across different classes. The model is further evaluated on a new subset composed of 250 manually annotated sentences. Out of these 250 sentences, 29 are wrongly predicted, representing an error rate of 11.6%. Table 6 provides the values for Accuracy, F1-score, Precision, and Recall. Finally, the complete dataset is fed into the model to classify the 65,289 relations. Figure 5 illustrates the final distribution of the relationships, where "Work/Study" is the most frequent class (40.3%), and "Kinship" is the least frequent (12.4%).

5 Results and Conclusion

The evaluation of the NER model demonstrates its growing capacity to accurately recognize entities within the provided text data, suggesting that the

Metrics	Values
Accuracy	0.884
Precision	0.90
Recall	0.884
F1-Score	0.888

Table 6: Evaluation of a manually annotated subset of 250 sentences

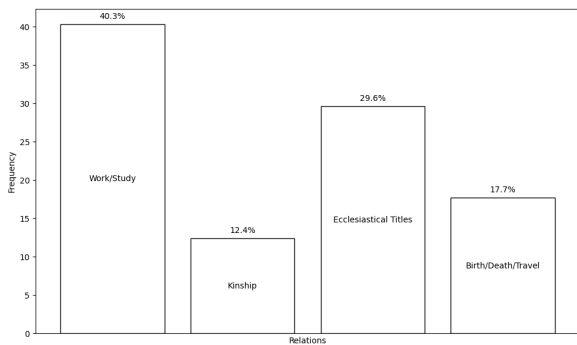


Figure 5: The distribution of all extracted relations across classes in the final and complete dataset

model effectively fine-tunes its parameters to better capture these entities, leading to significantly improved performance. Similarly, the results from the evaluation of the Relation Classification model reveal consistent and progressive enhancements throughout the training process. Beginning with a moderate initial accuracy of 61.8%, the model exhibits significant improvements over time. By the 20th epoch, it achieves a loss of 0.2037 and an accuracy of approximately 92.1%, indicating substantial progress from the initial stages. The declining loss values reflect a reduction in prediction errors, while the positive trends in Recall, Precision, and F1-score demonstrate the model's increasing ability to correctly identify relationships across various classes. The steady improvement in these metrics highlights the model's enhanced capability to capture true positives while minimizing false positives and false negatives. The final evaluation on a new subset of 250 manually annotated sentences, which resulted in an error rate of 11.6%, further supports the model's effectiveness in classification tasks. The high overall accuracy and balanced metrics indicate robust performance. Figure 6 shows that the model excelled in classifying "Work/Study" and "Ecclesiastical Title," achieving high true positive counts of 111 and 60, respectively. However, there are notable misclassifications, particularly in the "Birth/Death/Travel"

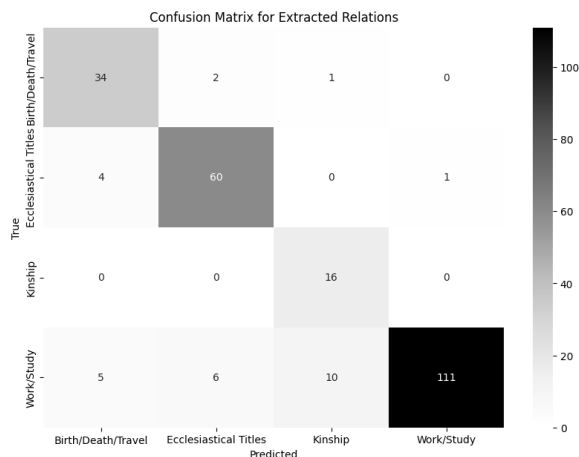


Figure 6: Confusion Matrix for the Extracted Relations

and "Kinship" categories. These issues may be attributed to the class imbalances highlighted in Figure 5, suggesting areas for improvement in distinguishing these categories. Finally, the classification of 65,289 relations within the complete dataset reveals a meaningful distribution of relationships, with "Work/Study" as the most frequent and "Kinship" as the least common. These findings not only underline the model's ability to accurately recognize and categorize different relationships within the text data, but also align with the patterns that are evident from simply reading the texts, suggesting that the model is well-optimized for the task at hand.

Cultural heritage has a huge role in shaping national and cultural identity, promoting tourism, and attracting visitors; on the other hand, digital technology enables cultural institutions to offer global access to their collections, including items rarely exhibited due to space constraints or fragility (Sporleder, 2010). Therefore, in the era of rapid digital transformation where the volume of data has surged dramatically, the challenge is not merely increasing the number of information but transforming it into valuable insights. Whereas digitalization has initiated a new era, the true potential of data can only be unlocked through innovative methodologies that transcend conventional approaches and offer new opportunities for exploration. Within this context, the endeavor to employ new tools holds profound significance, because whereas traditional methodologies, reliant on manual studies of literature and documents by academics, have laid the groundwork, they may not have fully revealed all the potentiality of knowledge. This work wants to be an opportunity to bridge this gap, leveraging nat-

ural language methodologies to delve deeper into these big collections of data. The present study proposes a two-step procedure to extract entities and relations from text data concerning ecclesiastical cultural heritage. A custom pipeline for Named Entity Recognition and a multi-classification model are exploited with manually annotated data, to extract and classify 65,289 relations. A differentiated evaluation process is performed to assess the significance of the proposed methodology. For future work, we intend to train the model using a larger corpus of annotated data to enhance the classification performance. Finally, the ultimate goal of this study is to transform unstructured text into structured data to create a graph database, specifically a Knowledge Graph. This Knowledge Graph will serve as a valuable tool for operators studying extensive texts, enabling them to uncover new connections that may not have previously emerged. Additionally, the extracted information will be integrated into a recommender system designed to assist users on the BeWeb platform. BeWeb is a website that provides access to a vast amount of data concerning ecclesiastical cultural heritage. By offering personalized suggestions, the recommender system aims to enhance user engagement and foster the creation of communities around shared interests.

6 Limitations

This research presents some limitations that can be viewed as opportunities for future studies. One key aspect is the number of annotators involved in the annotation process. Increasing the number of annotators could lead to more standardized annotations and, ultimately, more accurate predictions, thereby enhancing the overall quality of the research. Moreover, as suggested by the reviewers, it would be interesting to explore the differences in predictions between sentences containing only one entity and those with multiple entities, possibly by incorporating entity name embeddings. This analysis could provide valuable insights into how the model handles different situations. This research direction has the potential to yield meaningful results, further increasing the accuracy and reliability of the described system. Finally, it is important to emphasize that this research is focused solely on the Italian language; however, a similar approach could be implemented for texts of other languages.

Acknowledgments

This work was sponsored by the Italian Research Program PON (Programma Operativo Nazionale) “Ricerca e Innovazione” 2014-2020 - Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione” - Università degli Studi di Messina - Dottorato di Ricerca in “Economics, Management and Statistics” - XXXVII ciclo. This study was supported by IDS&Unitelm; a special thank goes to Angelo Cingari, Maria Teresa Rizzo, Massimo Currò, Nuccio Castorina, Silvia Tichetti, and finally the supervisor of the project Professor Edoardo Otranto. Gratitude is extended to the reviewers for their time and effort in reviewing the manuscript; their valuable comments and suggestions have significantly contributed to enhancing its quality.

References

- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, and X. ... Zheng. 2016. [TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems](#). *Preprint*, arXiv:1603.04467.
- D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil. 2018. [Universal sentence encoder for English](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.
- Q. Chen, M. El-Mennaoui, A. Fosset, A. Rebei, H. Cao, C. O’Beirne, S. Shevchenko, and M. Rosenbaum. 2022. [Towards mapping the contemporary art world with artlm: an art-specific nlp model](#).
- J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- D. Gonzalez and L. Andrew. 2014. [Rethinking recommendations: Digital tools for art discovery](#).
- N. Kambhatla. 2004. [Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations](#). In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions, ACLdemo ’04*, page 22–es, USA. Association for Computational Linguistics.

- D. P. Kingma and J. Ba. 2014. [Adam: A method for stochastic optimization](#). *CoRR*, abs/1412.6980.
- S. Russo. 2014. [Beweb. the cross portal of cultural ecclesiastical heritage](#). *JLIS.it*, 5(2):147–157.
- C. Santini, M. Tan, O. Bruns, T. Tietz, E. Posthumus, and H. Sack. 2022. [Knowledge extraction for art history: the case of vasari’s the lives of the artists \(1568\)](#).
- C. Sporleder. 2010. [Natural language processing for cultural heritage domains](#). *Language and Linguistics Compass*, 4(9):750–768.
- Fabian Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. [Combining linguistic and statistical analysis to extract relations from web documents](#). In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 712–717.
- P. G. Weston, F. D’Agnelli, S. Tichetti, C. Guerrieri, and M. T. Rizzo. 2017. [Authority data and cross-domain intersection within aggregation portals. the case of beweb](#). *JLIS.it*, 8(1):139–154.
- I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.
- D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. 2014. [Relation classification via convolutional deep neural network](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.
- P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. 2016. [Attention-based bidirectional long short-term memory networks for relation classification](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience

Lucija Krušić

Department of Digital Humanities

University of Graz

lucija.krusic@uni-graz.at

Abstract

This study presents the development of a sentiment-annotated corpus of historical newspaper texts in Austrian German, addressing a gap in annotated corpora for Natural Language Processing in the field of Digital Humanities. Three annotators categorised 1005 sentences from two 19th-century periodicals into four sentiment categories: positive, negative, neutral, and mixed. The annotators, Masters and PhD students in Linguistics and Digital Humanities, are considered semi-experts and have received substantial training during this annotation study. Three tools were used and compared in the annotation process: Google Sheets, Google Forms and Doccano, and resulted in a gold standard corpus. The analysis revealed a fair to moderate inter-rater agreement (Fleiss' kappa = 0.405) and an average percentage agreement of 45.7% for full consensus and 92.5% for majority vote. As majority vote is needed for the creation of a gold standard corpus, these results are considered sufficient, and the annotations reliable. The study also introduced comprehensive guidelines for sentiment annotation, which were essential to overcome the challenges posed by historical language and context. The annotators' experience was assessed through a combination of standardised usability tests (NASA-TLX and UEQ-S) and a detailed custom-made user experience questionnaire, which provided qualitative insights into the difficulties and usability of the tools used. The questionnaire is an additional resource that can be used to assess usability and user experience assessments in future annotation studies. The findings demonstrate the effectiveness of semi-expert annotators and dedicated tools in producing reliable annotations and provide valuable resources, including the annotated corpus, guidelines, and a user experience questionnaire, for future sentiment analysis and annotation of Austrian historical texts. The sentiment-annotated corpus will be used as the gold standard for fine-tuning and evaluating machine learning models for senti-

ment analysis of Austrian historical newspapers with the topic of migration and minorities in a subsequent study.

1 Introduction

Sentiment analysis (SA), the automatic identification of attitudes, opinions, and emotions in textual data, has been popular since the early 2000s (Liu, 2012). Deriving from Natural Language Processing (NLP), it was initially used to study contemporary data, including reviews and microblog posts. Recently, it has gained prominence in Digital Humanities (DH), expanding beyond contemporary texts to historical and literary texts (Häußler and Gius, 2023; Koncar et al., 2020; Kim and Klinger, 2019). However, texts such as newspaper articles, novels, letters, and poetry, which are commonly studied in DH, pose a challenge due to their formal structures and historical nuances, making sentiment analysis difficult (Kaur and R. Saini, 2014).

Traditional dictionary-based SA methods, heavily relied upon in DH, involve annotating words and phrases with sentiment values. Although this method is easily interpretable and transparent, sentiment dictionaries suffer from low reusability and do not consider word context, missing nuances such as sarcasm or negation (Schmidt et al., 2021c; Schmidt and Burghardt, 2018). To address these limitations, context-sensitive transformer-based machine learning models such as BERT have been developed (Devlin et al., 2018) (Suissa et al., 2022). These models require less annotated data than traditional ML algorithms (such as BOW or TF-IDF), since they can be pre-trained on large, unannotated datasets. This means that pre-training is usually done once and the model can then be further fine-tuned for various specific purposes and tasks, such as e.g. named-entity recognition or sentiment analysis, using a smaller annotated corpus (of e.g. sentences, plays or verses with corresponding sentiment annotation). This is particularly use-

ful in DH, where the annotation process is often tedious and time-consuming due to the complexity of literary and historical texts.

However, good quality annotations are crucial for the accuracy of the models with which they are fine-tuned. To better understand the conditions necessary for creating a high-quality sentiment-annotated fine-tuning corpus for texts in DH, recent studies have focused on the annotation process (Al-Laith et al., 2024; Sprugnoli et al., 2023), annotator behavior (Schmidt et al., 2018), and annotation tools (Schmidt et al., 2019). These studies discuss the relevance of expert versus non-expert annotators, the optimal tool for sentiment annotation, and the importance of guidelines. These insights informed the approach taken in this study, as will be shown in the next chapters.

There is still a gap in sentiment-annotated DH corpora that could be used for fine-tuning Machine Learning models, one such model being presented in Schweter (2020). This model was trained on non-annotated historical newspapers and offers the possibility of further fine-tuning with an annotated corpus, for a specific task - such as named-entity recognition or sentiment analysis. However, newspaper texts pose various annotation challenges: historical language and context, discriminatory language, sarcasm and metaphors.

The following sections present the current state of sentiment annotation in Digital Humanities and describe the creation of a sentiment-annotated corpus of Austrian historical newspapers through an annotation study. The annotations, the annotation process and annotation tools are evaluated quantitatively and qualitatively. This evaluation identifies key challenges and provides guidelines for annotating historical newspaper texts. These initial results will guide future refinements of the corpus, which will be openly accessible on Zenodo in accordance with the FAIR principles

2 Sentiment annotation for the Digital Humanities

Sentiment analysis (SA), in the context of Digital Humanities (DH), has often been used to answer specific research questions related to literary or historical studies. It has been used to analyse German plays (Schmidt et al., 2021a), Spanish Enlightenment periodicals (Koncar et al., 2021), Spanish song lyrics (Hernández-Lorenzo et al., 2022) and conflict in German novels (Häußler and Gius,

2023). Texts investigated by DH, such as newspapers, novels, poetry, and drama, present unique challenges due to their formal structures and historical linguistic nuances, making SA and annotation particularly complex (Kaur and R. Saini, 2014).

Sprugnoli et al. (2023) lists several aspects of the annotation process that are to be considered when constructing an annotation study: classification granularity, type of annotator (expert, non-expert or crowd workers), perspective, unit of annotation, and language of annotation unit. A further consideration (Schmidt et al., 2019) is the choice of annotation collection tool, which can have an effect on the annotation experience. These factors can impact the inter-rater agreement, which informs about the quality of annotations and is crucial in the development of a gold standard corpus.

In SA, two primary classification tasks are typically addressed: polarity and emotion analysis. Polarity analysis focuses on determining the direction of the sentiment within the text, often classified into categories such as positive, negative, and neutral (Liu, 2012). For more complex analyses, polarity classification may involve finer distinctions, such as differentiating between highly negative and highly positive sentiments, often using a numerical scale or additional categories. Sprugnoli et al. (2023) annotate four categories: positive, negative, neutral, and mixed. On the other hand, emotion analysis refers to the classification into emotional categories, often following Ekman's theory of basic emotions (Ekman, 1992) or Russell's circumplex model (Russell, 1980). Schmidt et al. (2019) conduct a polarity annotation study of German historical plays by G.E. Lessing, using extended polarity categories - negative, positive, neutral, mixed, uncertain, and other. In a subsequent study (Schmidt et al., 2021b), they conducted an additional annotation study with emotion categories. In these studies, inter-rater agreement, measured by statistical measures of Cohen's or Fleiss' kappa and Krippendorff's alpha, decreases with the higher number of categories that need to be annotated (Sprugnoli and Redaelli, 2024). However, these studies show that the agreement on literary and historical texts ranges from poor to moderate agreement, due to subjectivity and difficulty of the annotation process.

When conducting sentiment annotation, experts are preferred annotators, due to their accuracy and deep understanding of complex texts (Sprugnoli et al., 2023). But, they are scarce and expensive

(Schmidt et al., 2018). Semi-experts, such as advanced students, provide a more accessible alternative with reasonably reliable results (Yeruva et al., 2020; Schmidt and Burghardt, 2018), while non-experts (e.g. obtained through crowd-sourcing), though less accurate, can be effectively utilised in large-scale projects with appropriate guidance and annotation schemes (Schmidt et al., 2018)).

Sentiment annotation can focus on two perspectives: the emotions the author intended to convey, or the emotions perceived by the reader (Sprugnoli and Redaelli, 2024). Most studies focus on sentiments as intended by the author of the text (Sprugnoli et al., 2023; Häußler and Gius, 2023; Schmidt et al., 2019), as the annotation from the perspective of the reader can lead to low inter-rater agreement, due to subjectivity of the task.

The unit of annotation is also significant, and highly dependent on the type of text. Sprugnoli et al. (2023) and Häußler and Gius (2023) annotate sentences, while Schmidt et al. (2018) annotate speeches in a larger play. Annotating a shorter unit can be beneficial, as it minimises the change in sentiment shifting within the annotation unit.

Traditionally, spreadsheets and Word have often been used as the main tool for annotation collection (Sprugnoli and Redaelli, 2024; Sprugnoli et al., 2023; Schmidt et al., 2018). Schmidt et al. (2019) compare various annotation tools, such as Word, WebAnno, CATMA, eMargin and Sentimentator. They report using Sentimentator (a dedicated annotation tool) and Word increases annotator levels of certainty, thus, making the choice of annotation tool important for obtaining high-quality annotations. They employ standard usability and user experience questionnaires, NASA-TLX (Hart and Staveland, 1988) and User Experience Questionnaire (UEQ-S) (Hinderks et al., 2018), to assess user experience and perceived annotator workload.

With respect to the previous work presented in this section, the annotators in this study are semi-experts who have received extensive training in sentiment annotation. They annotate sentences in four categories: positive, negative, neutral and mixed, focusing on the sentiment intended by the writer. Furthermore, Google Forms (an online survey tool), Google Sheets (an online spreadsheet tool) and Doccano (Hiroki et al., 2018) are compared for the annotation process in order to establish an optimal tool for future annotation processes. To assess the usability of different annotation tools, Google

Forms, Google Sheets, and Doccano (Hiroki et al., 2018) were compared. The findings from this comparison, along with insights into annotator experiences, will inform the choice of tools for future sentiment annotation projects. Additionally, user experience was assessed using a combination of NASA-TLX, UEQ-S, and custom questions targeting the specific challenges of annotating historical texts.

3 Aims and research questions

The main aim of this study is establishing the optimal conditions for sentiment annotation of Austrian historical newspapers, with the goal of creating a reliable gold standard corpus for fine-tuning of ML models for sentiment analysis. This study aims to answer the following questions:

RQ1 Is using semi-expert annotators appropriate for the task of annotation of historical newspapers?

RQ2 How does the historical language and context of the texts influence the annotation process?

RQ3 How do the annotators perceive the difficulty of the annotation task?

RQ4 Which tool is most optimal for sentiment annotation of historical newspapers?

4 Methods

4.1 Corpus

The corpus used for the annotation consisted of 1005 sentences from two Austrian periodicals, “Neue Freie Presse” and “Das Vaterland”. The newspapers were published between 1850 and 1900. The corpus was created using Dynamic Topic Modelling with BERTopic (Grootendorst, 2022)), and through this process was automatically annotated with topics such as “migration”, “labour”, “Jews”, “Croats”, “Czechs”, etc. Sentences were used as the unit of annotation, with an average sentence length of 35.7 tokens, the shortest sentence having four tokens and the longest having 350 tokens. A sentence was used as the annotation unit because sentiment often changes within an article and sometimes even within a sentence.

4.2 Annotation process

The corpus was annotated by three semi-expert annotators (Masters and PhD students in Linguistics and Digital Humanities), two native German speakers and one fluent German speaker. The annotators were previously familiar with the task of sentiment

analysis and received additional training for the task of sentiment annotation. Each annotator was individually introduced to the corpus and the annotation process, followed by practical examples. The annotators were assigned to the annotation tasks for 3 months, 5 hours per week.

They individually annotated the sentiment in four categories:

- Positive (positive sentiment is expressed in the sentence)
- Negative (the sentence expresses a negative sentiment)
- Neutral (there is no sentiment in the sentence)
- Mixed (two sentiments are expressed, it is not possible to find a clear dominant one).

With regard to annotation perspective (Sprugnoli et al., 2023) they annotated the sentiment the author intended to convey through the sentence. The process was organised in stages, with group meetings after each round of annotation to exchange feedback, provide further training and resolve any uncertainties. In the first stage, 50 comments were provided in the form of a Google Forms survey with multiple choice questions. No additional information was provided in this round. In the second stage, 232 annotation units were provided to the annotators via a spreadsheet in Google Sheets. This time, in response to annotator feedback, the previous and subsequent sentences were provided as additional context, as well as the name of the journal and the date of publication. A column for comments was also added so that annotators could leave comments about their annotation choices if they felt it was necessary.

In the following 5 annotation rounds, they were given the remaining 723 sentences, divided into separate annotation tasks. The sentences were annotated using Doccano, an open source data labelling tool for machine learning tasks such as classification (Hiroki et al., 2018). Doccano was built in Python using the Django library, and an instance of it was deployed using Heroku for this annotation study. Doccano allows the upload and download of datasets in various non-proprietary formats (including csv). The main benefits are the ease of assigning annotation units to users, the ability to view one annotation unit at a time and navigate between them, and the ability to view additional information

about the annotation unit on the side of the screen. You can also easily track your progress, adding a gamification aspect to the annotation process. This has previously been shown to be beneficial to the user experience (Schmidt et al., 2019).

4.3 Evaluation

The annotations were evaluated for inter-rater agreement using Fleiss' Kappa (McHugh, 2012) and average percentage agreement (APA) overall and per category.

At the end of all annotation tasks, annotators were asked to complete a questionnaire (administered via Google Forms) about various aspects of the annotation process and the annotation tools used. The questionnaire comprised seven sections and a total of 25 questions. The complete questionnaire is available in Appendix A. The questionnaire included questions on the overall perceived difficulty of the annotation task, the perceived time taken to complete the annotations, and confidence in the annotations. The impact of historical language, context and specific linguistic features (such as sarcasm and metaphor) on the complexity of the annotation process was also examined. A section dedicated to the comparison of annotation tasks evaluated the ease of use of each tool and the speed of adaptation to the tool using a 5-point Likert scale.

In addition, following the recommendation of (Schmidt et al., 2019), two standardised usability tests were used to quantify the overall usability and user experience, namely the Nasa Task Load Index (NASA-TLX) (Hart and Staveland, 1988) and the User Experience Questionnaire (UEQ-S) (Hinderks et al., 2018). The NASA-TLX assesses the perceived workload of a task across different dimensions such as mental, physical and temporal (Schmidt et al., 2019). The scores are then combined and averaged into an overall workload score. The UEQ-S (User Experience Questionnaire - Short Version) (Hinderks et al., 2018) is used to quantitatively assess user experience across two key dimensions: Pragmatic Quality (PQ) and Hedonic Quality (HQ). PQ evaluates the usability and functionality of a product or task, indicating how easy it is for users to accomplish their goals. HQ measures the emotional appeal, reflecting how enjoyable, engaging, and motivating the product or task is for users. By analysing both PQ and HQ scores, the UEQ-S provides a comprehensive

overview of the user experience. Using the above metrics and questionnaire, the quality of annotation and the experience of annotators will be assessed.

5 Annotation evaluation

The inter-rater agreement, as measured by Fleiss' kappa, indicates a fair to moderate level of agreement (0.405), which is in line with expectations. The Average Percentage Agreement (APA) is 45.7% when all three annotators concur. However, for a sentence to be included in the gold standard, a majority vote – where two out of three annotators agree – is deemed sufficient. This resulted in a 92.5% agreement among at least two annotators across all annotation items in our corpus, which was used to create a gold standard of 930 items. Of the total number of annotations, 447 were classified as negative, 345 as neutral, 81 as positive, and 56 as mixed. The results indicate that the token length may have an impact on the annotations. Items rated as 'mixed' had a mean of 53 tokens, compared to means of 31, 29, and 38 tokens for 'positive', 'neutral', and 'negative' annotations, respectively. This indicates that additional splitting of the annotation items may be required to eliminate ambiguity. Notably, two annotators classified approximately half of the units as negative, while the third annotator rated 34% as negative and 45% as neutral, indicating that the most predominant sentiments were negative and neutral (see Table 1).¹

By closely examining the sentences in the annotation study, we can identify the causes of disagreement. For instance, the sentence “In den Beziehungen zwischen Polen und Czechen ist, trotz der vielen gegenseitigen Versicherungen brüderlicher Freundschaft, in jüngster Zeit — wie fast regelmäßig vor jedem Wiederzusammentritte des Reichsrathes — eine Spannung eingetreten.” (English translation: “In the relations between Poles and Czechs, despite the many mutual assurances of brotherly friendship, a tension has recently arisen—almost always before each reconvening of the Reichsrat.”) is highly ambiguous. It conveys both positive and negative sentiments between the subjects (Czechs and Poles) and suggests a possible disdain by the author (“almost always”). In contrast, sentences with complete agreement, such as “Treibt auch Noth einen Serben oder Walachen zur Arbeit, so strengt er sich durchaus nicht an.” (English translation: “Even if

necessity drives a Serb or Wallachian to work, they do not exert themselves at all.”), present a clearer sentiment and a more evident object of that sentiment.

6 Evaluation of annotator experience and annotation tools

The questionnaire yielded valuable insights into the attitudes and perceptions of the annotators regarding the annotation process. In terms of the complexity of the text annotation process, two participants rated it as "challenging" (4 on a 5-point scale), while one rated it as "moderate" (3 on a 5-point scale). The estimated time required for one annotation unit is between one and four minutes, with one annotation round taking between three and over five hours. Two annotators reported needing to take regular breaks from the annotation process. However, two out of three annotators report having high confidence (4 out of 5 on a Likert scale) in their annotations, while one reports moderate confidence (3 out of 5). In terms of specific difficulties, annotating longer texts was identified as particularly challenging, particularly when sentences were complex and required close attention. Another challenge identified was the need to remain objective, as one participant mentioned the difficulty of not letting personal beliefs influence the annotation process. The historical context and language of the texts had a notable impact on the annotation process. The participants indicated that the historical context affected their ability to annotate, with one annotator finding it particularly challenging. This emphasizes the importance of familiarity with the historical background when conducting sentiment analysis.

Moreover, all participants felt that the historical context significantly influenced their ability to annotate sentiment, reporting the need to independently research the historical background using resources like Wikipedia, the University Library Catalogue, Britannica, and the ANNO repository. The historical language, including vocabulary, grammar, and phrasing, also posed challenges, similarly influencing their ability to annotate sentiment.

Participants indicated that the clarity of what should be annotated—whether it was the sentiment of the language, sentiment towards a group, or the emotional state of the speaker—was not always clear. They reported that regular discussions and feedback sessions were useful for overcoming these

¹A Jupyter Notebook outlining the evaluation can be found at <https://github.com/lucijakrusic/SentiAnno/>

Annotation	Annotator 1	Annotator 2	Annotator 3
negative	550 (54.7%)	435 (43.33%)	346 (34.36%)
neutral	236 (23.51%)	423 (42.13%)	453 (45.12%)
positive	147 (14.61%)	73 (7.27%)	79 (7.87%)
mixed	71 (7.07%)	73 (7.27%)	126 (12.55%)

Table 1: Annotation results across three annotators.

challenges.

The use of sarcasm and metaphors in the texts presented a challenge for most participants. Two respondents found metaphors challenging, while one was affected by sarcasm, indicating a need for additional training or guidelines on handling figurative language in sentiment analysis.

Access to the previous and following sentences (context) was generally seen as helpful, with all participants agreeing that it aided in making more accurate annotations. However, opinions were divided on whether more context was necessary, with one participant suggesting that additional context could clarify ambiguous sentiments.

The overall NASA-TLX score was calculated by averaging the scores on six dimensions: mental, physical, temporal demand, performance, effort and frustration level. This equates to a score of 3.09 out of 5, indicating that the perceived workload is slightly above neutral. The annotators indicated that the level of workload was moderate overall, with some dimensions (such as mental demand, effort, and performance) rated higher than others. In terms of the UEQ-S, the Pragmatic Quality (PQ) achieved a score of 4.97 on a scale from 1 to 7 (1 reflects a poor experience in terms of usability, and a higher score of 7 reflects a good experience). This indicates that the task was moderately usable and clear. The hedonic quality (HQ) was rated at 4.45 out of 7 (with 1 indicating a less enjoyable experience and 7 indicating an emotionally satisfying one), indicating that the task was perceived as somewhat enjoyable and not unpleasant.

The questionnaire also assessed the usability of different annotation tools, with participants evaluating Google Forms, Google Sheets, and Doccano. All three annotators identified Doccano as the most intuitive tool, citing its clear layout, ease of navigation, and effective display of context as key factors. Additionally, the ability to leave comments and track progress was identified as a valuable feature. However, both Google Forms and Google Sheets are also considered relatively straightforward to

use and easily adaptable (see Figures 1 and 2). The annotators found Google Sheets less practical for navigating between the annotation units and viewing the full sentences. This is reflected in Figure 1, where one annotator noted that Google Sheets were difficult to use.

Lastly, annotators provided constructive feedback on how to improve the annotation process. One suggestion was to standardise the token length of context provided for each sentence, as inconsistencies sometimes made interpretation difficult. Another recommendation was to allow annotators to "correct" incomplete sentences by adding parts from adjacent sentences. These insights will be implemented in the following rounds of annotation.

7 Discussion

In this study, a sentiment-annotated corpus of Austrian historical newspaper texts was developed, with three semi-expert annotators categorizing sentences into four sentiment classes: positive, negative, neutral, and mixed. The inter-rater agreement, measured using Fleiss' kappa, resulted in a score of 0.405, indicating fair to moderate agreement and reflecting the inherent challenges in annotating historical corpora. This aligns with previous research ((Sprugnoli et al., 2023; Schmidt et al., 2019, 2018)) and highlights the difficulty of classifying complex texts, particularly those with mixed sentiments. This category, indicating both positive and negative sentiments within a single sentence, was the most challenging due to its higher token count, and presumably as a result, higher content ambiguity. This finding underscores the complexity of historical texts, where sentiments can shift within the same sentence or be expressed through nuanced language, including sarcasm and metaphors, that is difficult to categorise definitively. The study also observed a notable imbalance in sentiment categories, with the majority of annotations marked as negative or neutral. This distribution mirrors the historical context of the periodicals, which frequently adopted a critical stance toward migration

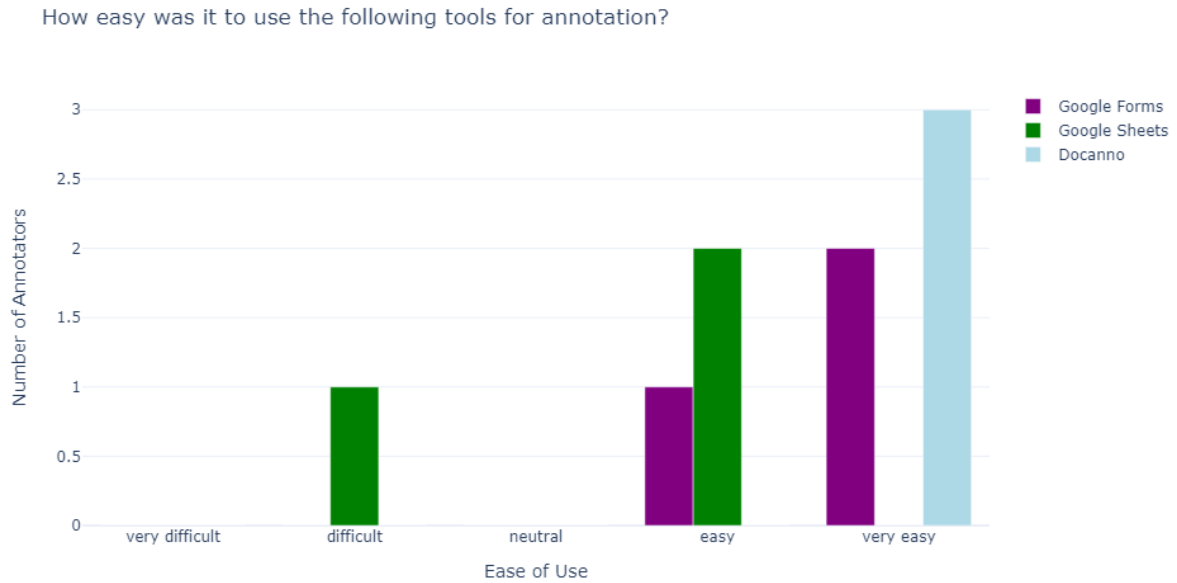


Figure 1: Annotation tool comparison on ease of use.

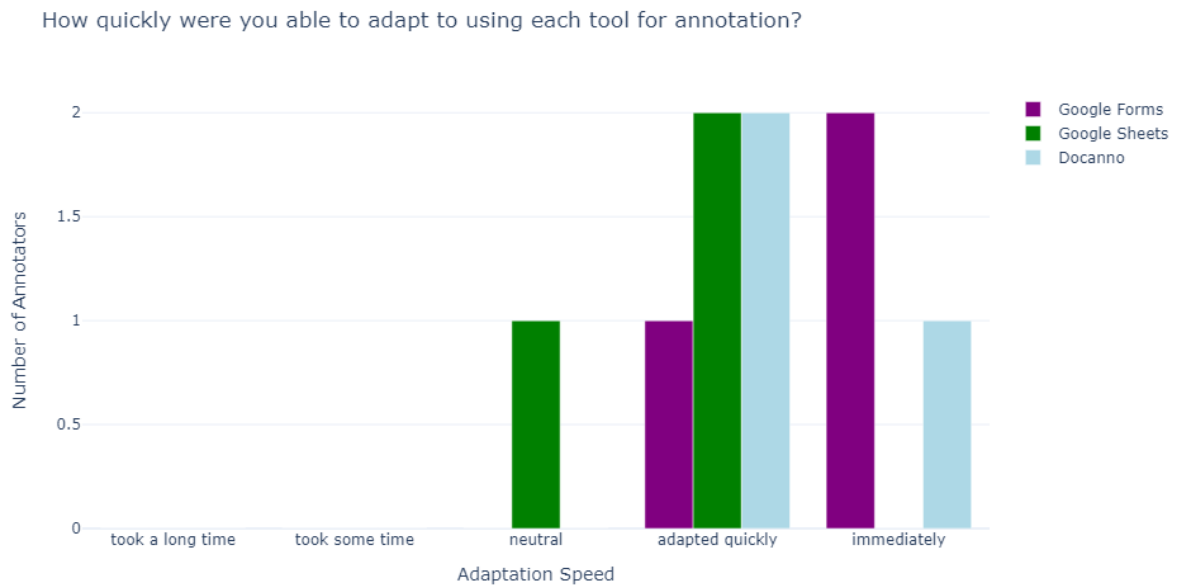


Figure 2: Annotation tool comparison on speed of adaptation.

and minorities. Despite these challenges, the annotation process achieved an Average Percentage Agreement (APA) of 45.7% for full agreement and 92.5% for majority vote (two out of three annotators). These results validate the reliability of the annotation process, allowing the creation of a gold standard corpus comprising 930 sentences, which will be extended in future annotation rounds.

Notably, the successful use of semi-expert an-

notators—advanced students—demonstrates that it is possible to achieve reliable annotations without relying on fully trained experts. This finding corroborates previous studies (Yeruva et al., 2020; Schmidt and Burghardt, 2018), reinforcing the notion that semi-experts can serve as an accessible yet effective alternative for similar tasks.

The annotation process was systematically supported by providing essential contextual informa-

tion, including references to the previous and following sentences, newspaper name, and date, which helped annotators interpret complex historical texts more accurately. Such contextual cues were particularly important given the nuanced language found in historical materials, where sentiments often shift within a sentence. The user experience questionnaire (Appendix A) revealed the cognitive and emotional demands on annotators, especially when dealing with historical language, figurative expressions like metaphors, and sarcasm. This feedback is valuable for improving future annotation workflows and provides a basis for comparing annotator experiences in similar tasks. Reusing the questionnaire, particularly its sections on historical language and context, could further enhance the systematic evaluation of annotation processes within Digital Humanities projects.

The standardized usability assessments, NASA-TLX and UEQ-S, highlighted the need to consider both cognitive workload and user engagement when designing annotation tasks. The NASA-TLX results showed that while the task was manageable, it required significant cognitive effort, particularly for complex, sentiment-laden historical texts. This finding aligns with [Schmidt et al. \(2019\)](#) and highlights the importance of considering workload when designing annotation tasks, particularly for complex historical texts. The UEQ-S results reveal a clear process (Pragmatic Quality) but suggest the task could be more engaging (Hedonic Quality). While Doccano proved to be the most user-friendly tool, with a positive impact on annotator efficiency and accuracy, there is room for improvement in user experience, particularly regarding task engagement. These results reinforce the need for comprehensive guidelines and tool evaluations, as well as attention to annotator workload, to ensure efficient and accurate sentiment annotation in Digital Humanities.

8 Conclusion

This study contributes to the field of Digital Humanities by presenting the first sentiment-annotated corpus of Austrian historical newspaper texts in Austrian German. Through the collaboration of three semi-expert annotators, 930 sentences were annotated for sentiment using a carefully designed process supported by tools like Doccano, Google Sheets, and Google Forms. The fair-to-moderate inter-rater agreement (Fleiss' kappa of 0.405) reflects the challenges of annotating historical texts,

where sentiment is often complex and contextually dependent.

A contribution of this study is the user experience questionnaire, which were specifically designed to assess the cognitive and emotional challenges encountered during the annotation process. The bespoke sections of the questionnaire not only provided valuable insights for improving subsequent annotation rounds but also offer a reusable framework for evaluating annotator experiences in other historical and literary annotation projects.

Furthermore, this study highlights the feasibility of employing semi-expert annotators in sentiment annotation, achieving reliable results through thorough guidelines and iterative feedback. Standardized assessments of usability and user experience, combined with the custom questionnaire, provided critical insights into annotators' cognitive demands and areas where the task could be improved.

By making the corpus openly available, this research offers a valuable resource for further sentiment analysis in Austrian German, particularly on topics such as migration, minorities, and labor rights. The findings and methodology outlined here will serve as a basis for future annotation projects, contributing to more nuanced and accessible sentiment analysis in historical and literary contexts.

Limitations

It should be noted that this study is subject to a number of limitations. Firstly, the limited number of annotators may impact the representativeness of the findings, particularly in terms of inter-rater agreement.

Secondly, the imbalance in sentiment categories, with a predominance of negative and neutral annotations, may have had an impact on the overall results. This imbalance reflects the content of the newspapers, but it also presents a challenge for model training and evaluation, as models may be biased towards these more common categories. Further rounds of annotation will be added to the corpus in the future, with the aim of reducing this imbalance.

Thirdly, the historical context and language of the texts presented significant challenges to the annotators, who had to navigate complex sentences and cultural references that may not have been immediately apparent. While the annotators were semi-experts, additional training or the use of annotators with expertise in history or media studies

could help to overcome some of the challenges identified.

Furthermore, the annotators indicated that they would have benefited from additional context accompanying the annotation unit and the ability to correct over-split annotation units, which will be addressed at a future stage in the annotation process.

The process of annotation of further data from historical newspapers will continue (with an extension of the temporal coverage and the addition of other newspapers with different political leanings). These limitations can serve as lessons that can be applied in the future to improve the creation of the gold standard.

Ethics Statement

This study was conducted with careful consideration of ethical principles, particularly in relation to the sensitive nature of historical newspaper content. The annotation units in this study included topics such as migration, labour, and minorities, which are often associated with discriminatory language and sentiments. The annotators were instructed to approach these texts with sensitivity and objectivity, ensuring that their annotations reflect the sentiment expressed in the text rather than their personal beliefs or biases.

Moreover, the historical context of the texts was acknowledged as a potential source of bias, both in the content of the texts themselves and in the interpretation by the annotators. To mitigate this, the annotators were provided with extensive training and were encouraged to research the historical background of the texts using reputable sources.

The study also adhered to ethical guidelines regarding the use of human participants. The annotators were informed about the purpose of the study, the tasks they were required to perform, and the potential challenges they might face. Their participation was voluntary and paid.

The study recognizes the potential impact of creating a sentiment-annotated corpus for historical texts, particularly in terms of how these texts may be interpreted and used in future research. The authors are committed to ensuring that the corpus is used responsibly and that any findings derived from it are presented in a manner that respects the historical context and the individuals represented in the texts.

Acknowledgements

We would like to express our sincere gratitude to Melanie Frauendorfer, Clara Hochreiter and Leona Münzer for their diligence and hard work on the corpus annotation. Their insights and feedback are greatly appreciated.

References

- Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024. [Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#).
- Paul Ekman. 1992. [An argument for basic emotions](#). *Cognition and Emotion*, 6(3-4):169–200.
- Maarten Grootendorst. 2022. [BERTopic: Neural topic modeling with a class-based TF-IDF procedure](#). ArXiv:2203.05794 [cs].
- Sandra G. Hart and Lowell E. Staveland. 1988. [Development of NASA-TLX \(Task Load Index\): Results of Empirical and Theoretical Research](#). 52:139–183.
- Laura Hernández-Lorenzo, Aitor Diaz, Alvaro Perez, Salvador Ros, and Elena González-Blanco. 2022. [Exploring Spanish contemporary song lyrics through Digital Humanities methods: Some thematic and structural properties](#). *Digital Scholarship in the Humanities*, 37(3):738–746.
- Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2018. [A Benchmark for the Short Version of the User Experience Questionnaire](#). In *Proceedings of the 14th International Conference on Web Information Systems and Technologies*, pages 373–377, Seville, Spain. SCITEPRESS - Science and Technology Publications.
- Nakayama Hiroki, Kubo Takahiro, Kamura Junya, Taniguchi Yasufumi, and Liang Xu. 2018. [{doccano}: Text Annotation Tool for Human](#).
- Julian Häußler and Evelyn Gius. 2023. [Operationalizing and Measuring Conflict in German Novels](#). In *Proceedings of the Computational Humanities Research Conference (CHR 2023)*, volume 3558.
- Jasleen Kaur and Jatinderkumar R. Saini. 2014. [Emotion Detection and Sentiment Analysis in Text Corpus: A Differential Study with Informal and Formal Writing Styles](#). *International Journal of Computer Applications*, 101(9):1–9.

- Evgeny Kim and Roman Klinger. 2019. [A Survey on Sentiment and Emotion Analysis for Computational Literary Studies](#). *Zeitschrift für digitale Geisteswissenschaften*. ArXiv: 1808.03137.
- Philipp Koncar, Alexandra Fuchs, Elisabeth Hobisch, Bernhard C. Geiger, Martina Scholger, and Denis Helic. 2020. [Text sentiment in the Age of Enlightenment: an analysis of spectator periodicals](#). *Applied Network Science*, 5(1):33.
- Philipp Koncar, Bernhard C. Geiger, Christina Glatz, Elisabeth Hobisch, Sanja Sarić, Martina Scholger, Yvonne Völkl, and Denis Helic. 2021. [A Sentiment Analysis Tool Chain for 18th Century Periodicals: Experimente in den Digital Humanities](#).
- Bing Liu. 2012. *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers, Vermont, Australia.
- Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.
- James A. Russell. 1980. [A circumplex model of affect](#). *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Thomas Schmidt and Manuel Burghardt. 2018. [An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing](#). In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.
- Thomas Schmidt, Manuel Burghardt, and Katrin Dennerlein. 2018. [Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior](#). pages 47–52. Sofia, Bulgaria. Conference Name: Annotation in Digital Humanities (annDH) Meeting Name: Annotation in Digital Humanities (annDH).
- Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021a. [Emotion Classification in German Plays with Transformer-based Language Models Pre-trained on Historical and Contemporary Language](#). In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 67–79, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.
- Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021b. [Towards a Corpus of Historical German Plays with Emotion Annotations](#). pages 11 pages, 741719 bytes.
- Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021c. [Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays](#). *Fabrikation von Erkenntnis: Experimente in den Digital Humanities* -.
- Thomas Schmidt, Brigitte Winterl, Milena Maul, Alina Schark, Andrea Vlad, and Christian Wolff. 2019. [Inter-Rater Agreement and Usability: A Comparative Evaluation of Annotation Tools for Sentiment Annotation](#).
- Stefan Schweter. 2020. [Europeana BERT and ELECTRA models](#).
- Rachele Sprugnoli, Francesco Mambrini, Marco Passarotti, and Giovanni Moretti. 2023. [The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace](#). *Italian Journal of Computational Linguistics*, 9(1).
- Rachele Sprugnoli and Arianna Redaelli. 2024. [How to Annotate Emotions in Historical Italian Novels: A Case Study on I Promessi Sposi](#). In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 105–115, Torino, Italia. ELRA and ICCL.
- Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. 2022. [Text analysis using deep neural networks in digital humanities and information science](#). *Journal of the Association for Information Science and Technology*, 73(2):268–287.
- Vijaya Kumari Yeruva, Mayanka Chandrashekar, Yungyung Lee, Jeff Rydberg-Cox, Virginia Blanton, and Nathan A Oyler. 2020. [Interpretation of Sentiment Analysis with Human-in-the-Loop](#). In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3099–3108, Atlanta, GA, USA. IEEE.

A Appendix A - Annotation questionnaire

A.1 Section 1 - General experience

1. How would you describe the difficulty of annotation of texts? (5 point Likert scale, very easy - very difficult)
2. How much time did you approximately spend on an annotation unit (sentence)? (short answer)
3. How much time did you approximately spend on 150 annotation units - one annotation round? (short answer)
4. Did you need to take frequent breaks from annotation due to the difficulty of the task? (multiple choice - yes/no/other)
5. How would you describe your confidence in your annotations? (5 point Likert scale, not confident at all - really confident)
6. Provided for you were the texts from two Austrian newspapers - "Neue Freie Presse" (NFP)

and "Das Vaterland" (VTL). The name (or the abbreviation) was also provided. Was one of them more difficult to annotate, and if so, which one? (multiple choice - NFP/VTL/not sure/both were equal in difficulty)

7. What did you find most difficult about the annotation process? (short answer)

A.2 Section 2 - Nasa Task Load Index (Hart and Staveland, 1988)

Please rate your experience on the following aspects of the task:

1. How mentally demanding was the task? (5 point Likert scale, very low - very high)
2. How physically demanding was the task? (5 point Likert scale, very low - very high)
3. How hurried or rushed was the pace of the task? (5 point Likert scale, very low - very high)
4. How successful were you in accomplishing the task? (5 point Likert scale, very low - very high)
5. How hard did you have to work to accomplish your level of performance? (5 point Likert scale, very low - very high)
6. How insecure, discouraged, irritated, stressed and annoyed were you? (5 point Likert scale, very low - very high)

A.3 Section 3 - User Experience Questionnaire (UEQ-S) (Hinderks et al., 2018)

Please rate your experience on the following aspects of the task:

1. Annoying - enjoyable (7 point Likert scale, left extreme - right extreme)
2. Not understandable - understandable (7 point Likert scale, left extreme - right extreme)
3. Slow - fast (7 point Likert scale, left extreme - right extreme)
4. Unpleasant - pleasant (7 point Likert scale, left extreme - right extreme)
5. Complicated - easy (7 point Likert scale, left extreme - right extreme)

6. Boring - exciting (7 point Likert scale, left extreme - right extreme)

7. Demotivating - motivating (7 point Likert scale, left extreme - right extreme)

8. Difficult to learn - easy to learn (7 point Likert scale, left extreme - right extreme)

A.4 Section 4 - Historical Language and Context

1. How much did the historical context of the texts affect your ability to annotate the sentiment? (5 point Likert scale, not at all - significantly)
2. How much did the historical language (e.g., vocabulary, grammar, phrasing) of the texts affect your ability to annotate the sentiment? (5 point Likert scale, not at all - significantly)
3. Did you feel the need to investigate the historical background of the texts on your own? (multiple choice - yes/no)
4. If yes, which resources did you use for this research (please specify)? (short answer)
5. How clear was it what should be annotated: the sentiment of the language, the sentiment towards a person/group of people, the sentiment towards a subject or the emotional state of the speaker? (5 point Likert scale, completely unclear - very clear)
6. If the task was unclear, what would have/has helped you overcome it? (short answer)

A.5 Section 5 - Specific Language Properties

1. How much did the appearance of sarcasm in the texts affect your annotation? (5 point Likert scale, not at all - significantly)
2. How much did the appearance of metaphors in the texts affect your annotation? (5 point Likert scale, not at all - significantly)
3. How much did having access to the previous and following sentences (context) help you in making accurate annotations? (5 point Likert scale, not at all - significantly)
4. Do you believe more context is necessary? (multiple choice - yes/no/other)

A.6 Section 6 - Tool usability comparison

1. How easy was it to use the following tools for annotation?
 - (a) Google Forms (5 point Likert scale, very easy - very difficult)
 - (b) Google Sheets (5 point Likert scale, very easy - very difficult)
 - (c) Doccano (5 point Likert scale, very easy - very difficult)
 - (d) How quickly were you able to adapt to using each tool for annotation?
 - (e) Google Forms (5 point Likert scale, took a long time - immediately)
 - (f) Google Sheets (5 point Likert scale, took a long time - immediately)
 - (g) Doccano (5 point Likert scale, took a long time - immediately)
 - (h) Which tool did you find the most intuitive to use for annotation tasks? (multiple choice - Google Forms/ Google Sheets/Doccano)
2. Please shortly elaborate why (short answer)

A.7 Section 7 - Additional feedback

1. This section is for any additional observations and remarks. How would you improve the annotation setup? Do you have any additional feedback or advice on how to improve the annotation process? (long answer)

It Is a Truth Individually Acknowledged: Cross-references On Demand

Piper Vasicek

Brigham Young University
Provo, UT
piper.vasicek@byu.edu

Courtnei Byun

Brigham Young University
Provo, UT
courtnei.byun@gmail.com

Kevin Seppi

Brigham Young University
Provo, UT
kseppi@cs.byu.edu

Abstract

Cross-references link source passages of text to other passages that elucidate the source passage in some way and can deepen human understanding. Despite their usefulness, however, good cross-references are hard to find, and extensive sets of cross-references only exist for the few most highly studied books such as the Bible, for which scholars have been collecting cross-references for hundreds of years.

Therefore, we propose a new task: generate cross-references for user-selected text on demand. We define a metric, *coverage*, to evaluate task performance. We adapt several models to generate cross-references, including an Anchor Words topic model, SBERT Sentence-Transformers, and ChatGPT, and evaluate their *coverage* in both English and German on existing cross-reference datasets. While ChatGPT outperforms other models on these datasets, this is likely due to data contamination. We hand-evaluate performance on the well-known works of Jane Austen and a less-known science fiction series *Sons of the Starfarers* by Joe Vasicek, finding that ChatGPT does not perform as well on these works; sentence embeddings perform best. We experiment with newer LLMs and large context windows, and suggest that future work should focus on deploying cross-references on-demand with readers to determine their effectiveness in the wild.

1 Introduction

A cross-reference is a connection between a source passage of text and another passage with bearing on the source passage. A cross-reference may contextualize, define, reinforce, restate, or even rebut the source passage, but a good cross-reference always elucidates the source passage in some way.

Good cross-references are hard to find because it takes familiarity with the text as a whole and a focused search through the text to find connections. While scholar-created sets of cross-references are

rare, the concept of self-reference within a text is ubiquitous. Consider the following two conversations from the film, “The Emperor’s New Groove” (Dindal, 2000):

Yzma: Fired? What do you mean, fired?

Kuzco: Um, how else can I say it? You’re being let go. Your department is being downsized. You’re part of an outplacement. We’re going in a different direction. We’re not picking up your option. Take your pick. I got more.

<later>

Yzma: Just think of it as you’re being let go, that your life’s going in a different direction, that your body’s part of a permanent outplacement.

Kronk: Hey, that’s kind of like what he said to you when you got fired.

This type of cross-reference, known as a “callback” in screenwriting, reuses or paraphrases a previous line. This type of cross-reference is easy to identify, since its express purpose is to *be* identified by the audience and to recontextualize the earlier line, often for laughs.

Most cross-references are subtler and more difficult to find, especially when not intended as cross-references by the writer(s).¹ Subtle cross-references can be found in one of two ways:

1. While reading, the cross-referencer happens to remember another connected passage—like Kronk recognizing the callback in “The Emperor’s New Groove”
2. The cross-referencer performs a focused search through the text specifically looking for connections to a source passage. When applied to each passage in a text, this translates to a complexity of $O(n^2)$.

¹See Appendix A for examples of subtler cross-references.

For print books, all desired cross-references must be discovered prior to printing since they cannot be added later. Electronic texts make it possible to discover cross-references dynamically instead. We therefore propose a fundamentally new task: provide cross-references on demand directly to readers for any passage they select.

Our contributions are as follows:

- Define the task: cross-references on demand.
- Define a metric to evaluate the task.
- Adapt several models to accomplish the task.
- Evaluate the model performance on three works in two languages, English and German.

We discuss prior work with cross-references in Section 2, define the task, cross-references on demand, in Section 3, discuss our methodology in Section 4, and present results in Section 5.

2 Prior Work

Before the advent of computers, all cross-reference sets were necessarily compiled manually. The more than 500,000 biblical cross-references originally published as the Treasury of Scripture Knowledge around 1830, were collected by “many authors ... over centuries” (Morton, 2010).

In 1973, the Church of Jesus Christ of Latter-day Saints formed a committee to create a new edition of the Bible for use by their congregations. This committee and “hundreds of workers”—mostly volunteers—took six years to create a cross-reference set for the new edition, despite evaluating existing Bible cross-references and using concordance software (Anderson, 1979).

The labor-intensive nature of these manual cross-referencing projects highlights the reasons why scholarly cross-reference sets are so rare.

Lund et al. (2019) investigated reducing the cost to create a cross-reference set by using topic modeling to suggest cross-references and crowdsourcing to evaluate them. However, creating a set of cross-references will be labor-intensive no matter how much technology improves. Even the Qur’an, although revered, was not extensively cross-referenced until 2022, the culmination of a project that took a decade to complete despite access to modern technology (Sirry, 2022).

There are over 130,000,000 books in publication (Taycher, 2010). It would be impossible to create sets of cross-references for all of them, but it might be possible to cross-reference them on demand instead.

Source Passage: **“It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.”** – Pride and Prejudice, ch.1

ChatGPT: **“Happy families are all alike; every unhappy family is unhappy in its own way.”** – Anna Karenina, ch.1

Sentence Embeddings: **“But there certainly are not so many men of large fortune in the world as there are pretty women to deserve them.”** – Mansfield Park, ch.1

Topic Modeling: **“It must make you better satisfied that your other four are single.”** – Pride and Prejudice, ch.53

Figure 1: The best examples from the top five suggestions generated by our best-performing models on the first sentence of Pride and Prejudice. ChatGPT suggests first lines from various literary works.

3 Defining Cross-references On Demand

What would cross-references on-demand look like? A reader selects a desired source passage, and a model returns suggested cross-references; see Figure 1 for an example of what this might look like. The difficulty of this task lies in finding *good* cross-references and evaluating cross-reference quality.

A good cross-reference enhances the reader’s understanding of the source passage. However readers are not monolithic, so we cannot expect every suggested cross-reference to be a good cross-reference for every reader. Even scholar-produced sets—which should only contain good cross-references—contain cross-references that some readers find unhelpful. If experts cannot produce universally good cross-references, we do not expect models to do so. Instead, we consider a model successful if a reader finds a satisfactory number of good cross-references in a relatively small number of suggestions.

To encapsulate this concept, we define the following metric which we refer to as *coverage*:

$$C = \frac{\sum_{d=1}^n f(d)}{n} \quad (1)$$

$$f(d) = \begin{cases} 1, & \text{if } |\{x_{d1}, x_{d2}, \dots, x_{di}\} \cap G| \geq t \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where C is the coverage, d is a source passage; n is the total number of source passages for which we evaluate suggested cross-references; x is a suggested cross-reference from the model; i is the number of cross-references suggested, G is the set of good cross-references, and t is the number of cross-references required to satisfy the reader. It is similar to precision@k, but instead of calculating the ratio of good suggestions to total suggestions, we determine whether or not critical mass of good suggestions has been achieved.

Coverage can be calculated for the entire text, or, conveniently, for a sample of the text with the result extrapolated to the rest of the text. Using *coverage* we can now compare the performance of various models to determine which produce the most satisfactory results.

4 Methodology

We consider which models to apply to cross-reference generation in Section 4.1; we discuss the Datasets to which we will apply them in Section 4.2; and we discuss using *coverage* to evaluate model performance in Section 4.3.

4.1 Model Selection

While any number of models could be adapted to cross-references on demand, we choose three to represent them.

Since Lund et al. (2019) applies topic modeling to static cross-reference set creation with good effect, it is logical to adapt their models to our task—see Section 4.1.1. These models rank cross-reference suggestions using topical similarity.

Sentence embeddings are a more mainstream way of comparing semantic similarity; we therefore adapt them as well—see Section 4.1.3.

Finally, it is unclear whether semantic similarity is the most effective way to find good cross-references. The context of the passage—including context external to the work—may prove essential to finding good cross-references. Large language models (LLMs) such as ChatGPT have proven astonishingly good at performing many difficult language tasks (OpenAI, 2022), and ChatGPT has recently been incorporated into a Bible study tool with an option to suggest cross-references (Norton, 2024). Therefore, we experiment with ChatGPT—see Section 4.1.5.

4.1.1 Topic-based Models

We adapt variations of two different topic-modeling-based models from Lund et al. (2019) to use as cross-reference generators and create a third, unique model using randomized topic-words.

We adapt the most successful model from Lund et al. (2019) to serve as a baseline. This model is based on Tandem Anchors (Lund et al., 2017), an extension of the Anchor Words algorithm for topic modeling (Arora et al., 2013) and uses 3000 topics. In order to generate cross-references, the model is given the entire text split up into passages (verses for the Bible and sentences for the other two works). The topics are chosen by randomly selecting a number of passages from the text equal to the desired number of topics. For each of these passages, we take the harmonic mean of the vector representation of all the words in the passage and add a small epsilon of $1e^{-10}$ in each dimension to avoid zero weights. The topics and text are then processed using the Anchor Words algorithm to produce topical weights for each passage. This topic-weight vector for each passage is then compared to the topic-weight vector for the source passage and the most topically similar passages are suggested as cross-references. We adapt this model further by sweeping the space of topics to determine an optimum number—which has not been done previously. We refer to models based on Tandem Anchors with *tand_n*, where “n” represents the number of topics.

The Anchor Words algorithm chooses words as topics based on a variation of the Gram-Schmidt process (Arora et al., 2013). Lund et al. (2019) also employed this model, which performed on par with their Tandem Anchors model. We therefore also adapt this model to our task. Topics are chosen by representing words in a high-dimensional space and attempting to pick n words to use as topics that maximally span that space, very similar to a convex hull. We refer to variations on this model with *gram_n*, where “n” represents the number of topics chosen for the model, and we sweep the topic space since this has not been done previously.

Using 3,000 or more topics for *gram_n* may not be sensible. For perspective, our datasets have a vocabulary between 9,000 and 20,000 words depending on preprocessing, meaning 3,000 topics is 15-30% of the vocabulary. This was not the expected use of topic modeling or Anchor Words, and it is unclear how well the modified Gram-Schmidt

process for selecting topic words will work at such high saturation. We therefore employ a model that selects n words from the vocabulary at random, and adds a small epsilon of $1e^{-10}$ in each dimension. We refer to this model as *rand_n*, where “n” represents the number of topics chosen for the model.

4.1.2 Preprocessing

For topic-based models—which use a bag-of-words approach—we remove stopwords and employ a stemmer: the Porter stemmer (Porter, 1980) for all English datasets, and the Snowball German stemmer for *Bible-GER* (Porter, 2001).

4.1.3 Sentence Embedding

We employ SBERT sentence-transformer models for sentence embedding (Reimers and Gurevych, 2019, 2020). For English data, we use all-mpnet-base-v2 which currently has the best average performance among available SBERT models. For German data we use paraphrase-multilingual-mpnet-base-v2, which currently performs best among multilingual SBERT models. Similar to the topic-based algorithms, we use the embedding vectors to rank the similarity of each passage to the source passage and suggest the most similar passages.

4.1.4 Distance Metrics

Lund et al. (2019) explored several distance metrics and ultimately evaluated their models using cosine similarity. However, a close viewing of their metric comparison results suggests that cityblock distance performs on par with cosine similarity for the most similar passages evaluated. We therefore include both cosine similarity and cityblock/Manhattan distance to determine semantic similarity.

4.1.5 ChatGPT

Finally, we employ the GPT-3.5-turbo model using default hyperparameters (OpenAI, 2022) to generate cross-references with the following prompt:

I am reading [TEXT_NAME] and want to find some good cross-references for [REFERENCE_AND/OR_QUOTE]. Can you suggest some good cross-references?

replacing [TEXT_NAME] with the name of the book (e.g. the Bible) and replacing [REFERENCE_AND/OR_QUOTE] with a reference to the passage (only for Bible data), the text of the passage, or both (e.g. “Genesis 1:1 - In the beginning,

God created the heavens and the earth.”)². We evaluate the cross-references generated, ignoring other generated text. We evaluate the first generation produced for each passage.

4.2 Datasets

We evaluate our models on their ability to suggest good cross-references for three different texts. First, the Bible for which there are expansive, freely available cross-reference sets. Second, for the works of Jane Austen which are widely known and analyzed, but for which there are no existing cross-reference sets. Third, for a science fiction series, *The Sons of the Starfarers*, by Joe Vasicek, which is less well-known and unlikely to be included in ChatGPT’s training data.

4.2.1 The Bible

The Bible allows us to evaluate results at scale using existing cross-reference sets. “The Treasury of Scripture Knowledge, Enhanced” (TSKE) (Morton, 2010) is a cross-reference resource based on the original TSK. This set of cross-references is especially useful since it contains an impressive 670,000 cross-references and contains cross-references for 96% of verses. While not the most expansive cross-reference set—there is at least one Biblical cross-reference set that boasts over 900,000 cross-references (Smith, 2016), the TSKE is freely available for download.

OpenBible.info has cross-references seeded from the TSK and other open source cross-reference sets and allows users to upvote helpful cross-references and downvote unhelpful ones. We use the OpenBible.info cross references and attached up/downvotes. This yields multiple cross-reference sets, allowing us to simulate readers with different views of what is a good cross-reference. We use the set of cross-references from OpenBible.info that have at least as many upvotes as downvotes—and call this *Open*. We also use the set of cross-references that have a minimum of 5 net upvotes and call this *Open5*.

Beyond the existence of expansive cross-reference sets, the Bible is a useful text because of its many translations. At the time of this writing, there were 3,035 bible translations available in 2,014 languages on Bible.com. For English

²When prompting ChatGPT with the passage text without the reference, we change the last sentence to “Can you suggest some good *quotes* as cross-references?” Without adjusting it, ChatGPT tends to ask for a reference or describe a scene.

	tand_2000		tand_3000		tand_4000		tand_5000	
	cos	city	cos	city	cos	city	cos	city
TSKE	0.307	0.393	0.328	0.416	0.357	0.445	0.353	0.442
Open	0.231	0.299	0.248	0.320	0.272	0.345	0.270	0.340
Open5	0.141	0.190	0.146	0.202	0.161	0.216	0.158	0.211

Table 1: The results for the tandem anchors model using cosine and cityblock distance metrics. Tandem anchors with 4,000 topics performs best for each dataset using the cityblock distance metric.

	gram_2000		gram_3000		gram_4000		gram_5000		gram_10241	
	cos	city	cos	city	cos	city	cos	city	cos	city
TSKE	0.290	0.327	0.334	0.379	0.381	0.447	0.380	0.443	0.379	0.445
Open	0.217	0.248	0.252	0.293	0.289	0.348	0.289	0.345	0.288	0.346
Open5	0.108	0.127	0.135	0.160	0.163	0.204	0.161	0.206	0.162	0.203

Table 2: Results for the gram_n model using cosine and cityblock metrics. 4,000 topics performs best, except for the Open5 cross-reference set with the cityblock metric for which 5,000 topics performs best.

Bible experiments, we use the text of the English Standard Version (ESV) of the Bible, since that is the version of the Bible on OpenBible.info. In 2022, the Luther Bible was the bestselling German translation of the Bible according to the German Bible Society (Bigl, 2023). We evaluate the cross-reference sets using this translation of the Bible which we refer to as *Bible-GER*.

4.2.2 The Works of Jane Austen

We know good cross-references exist for the Bible, but it is unclear to what extent this is true for other texts. Also, data contamination, specifically the ubiquity of Bible-cross-references, almost certainly gives ChatGPT an edge when it comes to Bible data. Therefore, Jane Austen’s novels are a great dataset to explore ChatGPT’s ability to perform on a text with which it is very familiar but for which there are not existing cross-reference sets.

We ask our three best-performing models to suggest 5 cross-references each for 100 randomly selected sentences from the works of Jane Austen. Two of our authors then hand-evaluate each suggestion.

4.2.3 Sons of the Starfarers

We include *Sons of the Starfarers* by Joe Vasicek because it is data that ChatGPT has never seen. We experiment with both one-shot prompting, and fine-

tuning chatGPT on this text, to see what ChatGPT can do with an entirely unknown text.

4.3 Coverage

We report coverage at $n = 1$ and $i = 5$. In other words, a passage adds to the coverage if at least one good cross-reference appears in the first 5 suggestions. We assume that a typical reader, interested in the source passage, will be willing to read 5 suggested cross-references before giving up on finding a good one. Future research could be done with actual readers to validate these assumptions. Coverage results on Bible data for a range of values of n and i are given in Appendix B.

5 Results

We report results on the English Bible, followed by Bible-GER, followed by the works of Jane Austen.

5.1 Topic sweeps

We sweep the topic space with a low of 50 topics, increasing until we no longer see improvements in coverage. Performance gains steadily from 50 topics with best results around 4,000 topics. We report results for best-performing number of topics.

For the tand_n model, 4,000 topics always performs best. See Table 1 for results for the tand_n model surrounding 4,000 topics.

	rand_4000		rand_8000		rand_9000		rand_10241		rand_14730	
	cos	city	cos	city	cos	city	cos	city	cos	city
TSKE	0.363	0.421	0.443	0.506	0.461	0.529	0.465	0.547	0.469	0.546
Open	0.278	0.326	0.346	0.400	0.360	0.419	0.364	0.434	0.368	0.435
Open5	0.154	0.193	0.208	0.256	0.229	0.270	0.219	0.280	0.225	0.286

Table 3: Bible *coverage* results for the rand_n model using both cosine and cityblock metrics. This model performs best when given the entire vocabulary as topics, sometimes with stemmed vocabulary (rand_10241), sometimes with unstemmed vocabulary (rand_14730).

	embeddings		ChatGPT		
	cos	city	ref	quote	both
TSKE	0.564	0.562	0.50	0.60	0.66
Open	0.440	0.437	0.44	0.55	0.54
Open5	0.294	0.288	0.47	0.50	0.45

Table 4: Bible *coverage* results for the sentence embedding and ChatGPT models.

The gram_n model also tops out at 4,000 topics, except on *Open5*, where 5,000 topics using the cityblock distance has a very slight advantage, gaining 0.2% *coverage*. Interestingly, despite the fact that the tand_n model outperforms the gram_n model when creating a static set of cross-references in Lund et al. (2019), here the gram_n model outperforms the tand_n model in terms of coverage for nearly all numbers of topics. See Table 2 for the results for the topic sweeps surrounding 4,000 topics for the gram_n model.

Perhaps the most interesting of the topic-modeling-based results is for the rand_n model which continues to increase its performance beyond 4,000 topics, and in fact increases in *coverage* until we max out the vocabulary at rand_10241. After maxing out the stemmed vocabulary, we run the model without stemming. Maxing out the unstemmed vocabulary at gram_14730, we see very small gains in most instances. See Table 3 for the results for the rand_n topic sweeps.

Surprised by this outcome, we run gram_10241—the whole vocabulary—to ensure that gram_4,000 is not a local minimum. However, gram_10241 fails to improve on gram_4000, performing slightly worse in all cases. This is surprising since rand_10241 and gram_10241 use exactly the same

set of topic words at this point. The only difference between the two models is the epsilon of $1e^{-10}$ added in each dimension to the rand_10241 topics. We do not know why this epsilon produces such a large increase in coverage (7-10% depending on the cross-referencing set). The epsilon represents uncertainty about the exact position a particular topic word should occupy in the topic space. Perhaps this allows the model the leeway it needs to tie topic words to words they might otherwise ignore if they contained zero weights in some dimensions.

The best topic-based model is rand_14730.

5.2 Sentence Embeddings

The sentence embedding model outperforms rand_14730 by 0.5-1.4% using cosine similarity. See Table 4 for sentence embedding results.

5.3 Distance Metrics

Cityblock distance outperforms cosine similarity for all topic-modeling models by a wide margin—up to 9%. This is not too surprising since the cityblock metric (L1 norm) has been shown to outperform other norms in high-dimensional spaces (Aggarwal et al., 2001). Cosine similarity outperforms cityblock distance by a small margin for all sentence embedding models. This is likely due to the way each model represents data. In topic models each vector component represents a discrete topic. Sentence embedding vector components represent data more abstractly, with no single concept attached to a particular component.

5.4 ChatGPT

ChatGPT outperforms all other models we employ on the Bible dataset, achieving a performance 3.8-20.6% better than the next best performance. See Table 4 for the ChatGPT results. ChatGPT particularly outperforms other models on the Open5 set

	tand_3000 (cos)	rand_20984 (city)	embeddings (cos)	ChatGPT (quote+ref)
TSKE	0.270	0.490	0.432	0.588
Open	0.200	0.383	0.342	0.490
Open5	0.109	0.245	0.241	0.419

Table 5: Bible-GER *coverage* results for the baseline model (tand_3000), rand_20984 (the entire unstemmed German vocabulary), multilingual sentence embeddings, and ChatGPT given both the reference and passage text.

of cross-references. This is most likely because the Open5 cross-references tend to be very popular cross-references, and since ChatGPT almost certainly contains Bible cross-references in its training data, very popular cross-references are highly likely to reappear in ChatGPT-generated text.

5.5 German

For the Bible-GER dataset, ChatGPT also performs best, followed by rand_20984 (using the entire unstemmed German vocabulary), followed by the sentence embedding model, see Table 5. Each model performs worse on the German data than on the English data. Some of this may be because the cross-referencing set we are using to evaluate the data was compiled using English Bibles. Likely, it is also because of the language-specific nature of many of the models and tools we are employing, including the stemmer we use for preprocessing, the sentence embedding model, and ChatGPT.

We also saw one particular recurring error in the ChatGPT generations for the Bible-GER dataset. ChatGPT often merged the number in front of a Bible reference into the structure of the list, (e.g. “5. 5.1 Thessalonicher 2:3-4”, instead of “5. 1 Thessalonicher 2:3-4”). We do not see this behavior in English ChatGPT generations, but for the Bible-GER dataset it occurs in 37% of generations with an average of 2.3 errors in those generations.

5.6 Jane Austen

Our authors rated cross-references for Jane Austen wildly differently. However, sentence embedding widely outperformed rand_20948 (the full Jane Austen vocabulary), which widely outperformed by ChatGPT. See Appendix D for the numerical results for the Jane Austen data. Because ChatGPT performed so much worse for our Jane Austen data, we perform an error analysis of ChatGPT’s responses below.

5.6.1 ChatGPT Error Analysis

For 28 source passages, ChatGPT did not produce 5 total suggestions.

In 19 suggestions (4% of the time), ChatGPT gave recommendations for how to find cross-references for a particular sentence instead of suggesting cross-references (e.g. ‘Social Class and Morality: "Mansfield Park" explores themes of social class and morality. You can look for quotes that delve into the moral values and social hierarchies of the characters.’).

In 67 suggestions (14% of the time), ChatGPT recommended an entire scene instead of the explicitly requested quote (e.g. “Mr. Collins’s proposal to Elizabeth is a comical but cringe-worthy moment that relates to the theme of marriage and the importance of character in choosing a spouse.”)

Of the 377 quotes that ChatGPT suggested, 149 (40%) were not accurate quotes; 42 contained pieces of recognizable quotes and 107 appeared to be entirely fabricated (e.g. “Truth is always truth, either in the shape of a woman or a rhinoceros;”).

Of the remaining 227 verifiably accurate quotes, 47 (21%) were from other sources besides Jane Austen’s work including movies based on her novels and other literary works, and 142 (63%) had at least one duplicate among cross-reference suggestions for other sentences (e.g. “It is a truth universally acknowledged . . . ” was suggested 12 times). In other words, it seems that the more well-known a quote is, the more likely it is to be suggested as a cross-reference by ChatGPT, indicating that ChatGPT may not be cross-referencing related passages so much as suggesting popular quotes regardless of the context. This so-called “Matthew Effect” (Merton, 1968) of ChatGPT has been reported in citation generation and in environmental science (Salleh, 2023; Petiska, 2023). ChatGPT’s performance on this task suggests that unless the cross-referencing task has already been performed manually for a text, ChatGPT struggles

to find cross-references.

5.7 Sons of the Starfarers

As expected, when applying a one-shot prompt to *Sons of the Starfarers*, ChatGPT cannot suggest any real quotes from the novel. It either suggests quotes from other literature—often classic literature, which seems odd for the space genre. Or it suggests unattributed quotes that seem to be entirely fabricated.

After finetuning ChatGPT on a chapter completion task (see Appendix C for details), ChatGPT was still unable to generate any real quotes from the novel. The finetuned model also stopped suggesting quotes from other literary works and only suggested fabricated quotes. This may indicate that ChatGPT’s performance on Jane Austen’s works can be attributed to the prevalence of criticism and well-known quotes in ChatGPT’s training data.

6 Discussion

The big winner on Bible data is ChatGPT. However, ChatGPT has an unfair advantage over the other models in this use case. Specifically, ChatGPT has undoubtedly seen cross-references and discussion of cross-references in its training data. We should obviously use this unfair advantage to our benefit when possible (for a Bible study tool, ChatGPT may be the best option). However, we also need to consider how ChatGPT will perform on texts that do not already have extensive cross-references available (the majority of texts).

For the works of Jane Austen and *Sons of the Starfarers*, ChatGPT performs significantly worse than the semantic similarity methods. There are almost certainly some ways in which these problems could be mitigated, including using newer models or embedding the entire corpus of text in the prompt.

We do some preliminary experiments using large contexts with ChatGPT and OpenAI’s new gpt-4o model on quotes from *Sons of the Starfarers*. When including large portions of the text in the prompt both ChatGPT and gpt-4o were able to suggest quotes that were recognizably from the novel, with gpt-4o seemingly able to capture more nuanced connections. These methods may quickly become expensive (approximately \$0.25 per source passage when employing the widest possible context length) which could make cross-references on demand less widely available. The desired cross-referencing cor-

pus for a particular work may also be too large for even the largest context windows currently available, Google currently boasts a Gemini model with a 1 million token context (Pichai, 2024).³ However, as LLM models improve and costs come down, this may indeed be a viable solution. It may also be possible to implement some form of retrieval augmented generation (RAG) to achieve better results without needing to use exceedingly large context windows (Lewis et al., 2020). Indeed, we are hopeful that this task will be adopted broadly, and new models and methods will be adapted and developed to improve on our results. Meanwhile, for non-Biblical texts, we recommend using sentence embeddings for cross-references on demand.

7 Related Works

Our work is most similar to work on intertextuality, source attribution, and literary evidence retrieval. Forstall and Scheirer provide an in-depth description of the use of computational tools, including topic-modeling, to discover literary intertextuality. Source attribution has been long-studied, but recent work by Muther and Smith (2023) is similar to our work in that it uses language models to rank candidate text. The most similar task to ours is likely literary evidence retrieval as explored by Thai et al. (2022) who created a novel dataset for literary evidence retrieval to test the ability of models to match literary analysis with the quotation described by that analysis. Source attribution, intertextuality, and literary evidence retrieval all seek to find the source for a statement, working from the assumption that a ground-truth source exists, generally in another body of work. For cross-references, there is no assumed source. In stark contrast, the “ground-truth” is how useful the cross-reference is to the reader.

8 Conclusion

We proposed cross-references on demand, defined *coverage*—a metric to evaluate performance on this task, and showed the efficacy of three different models on producing cross-references for three texts in English and German.

ChatGPT outperforms other models on the Bible. However, it performs significantly worse on texts that do not have existing cross-references, includ-

³Some authors, such as Brandon Sanderson, routinely write hundreds of thousands of words per year (Sanderson, 2018).

ing those represented in its training data. Topic-modeling and sentence embedding models perform comparably on the Bible, but hand-evaluation of these models suggests sentence embedding performs significantly better. We suggest applying sentence-embedding models when implementing interactive cross-references for texts for which no cross-reference sets exist.

Further research could focus on validating model performance directly with readers as well as adapting other models to cross-reference on demand.

9 Limitations

This work is still exploratory, and as such has several limitations. First, and foremost is our heavy reliance on Bible data. We are largely restricted in the ability to evaluate texts at scale by the limited existence of other large-scale cross-referencing resources.

Second, while we did use ChatGPT to produce cross-references, many other LLM models are available including more sophisticated models. Future work could explore these and ways to improve large language model performance for cross-referencing. It may be that fine-tuning a model on a large dataset of cross-references from a variety of sources could yield better results. However, the lack of available cross-referencing resources outside the Bible could make this a difficult endeavor.

Thirdly, while *coverage* is a useful quantitative representation of overall reader satisfaction, it does not take into account more qualitative aspects such as the relevance or explanatory power or cross-references. Future work with users should both validate the *coverage* metric and explore qualitative attributes of the cross-references suggested when determining the success of an on-demand cross-reference system.

Finally, we limit ourselves to cross-references from within a single body of work. Future work could assess whether these approaches are as effective when texts from multiple sources are included in the corpus.

References

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pages 420–434. Springer.

Lavina Fielding Anderson. 1979. Church publishes first eds edition of the bible. *Ensign*, 10.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pages II–280.

Sven Bigl. [Bestseller luther-bibel: Nachfrage im jubiläumsjahr 2022 um ein drittel gestiegen](#) [online]. 2023.

Mark (director) Dindal. 2000. The emperor’s new groove. Film.

Christopher W Forstall and Walter J Scheirer. Quantitative intertextuality.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Emily Hales, and Kevin Seppi. 2019. Cross-referencing using fine-grained topic modeling. In *Proceedings of NAACL-HLT*, pages 3978–3987.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multi-word anchor approach for interactive topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 896–905.

Robert K Merton. 1968. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.

Timothy Morton. 2010. *Treasury of Scripture Knowledge, Enhanced*. BibleAnalyzer.com.

Ryan Muther and David Smith. 2023. [Citations as queries: Source attribution using language models as rerankers](#). *Preprint*, arXiv:2306.17322.

Oak Norton. 2024. [Introducing daniel, your scripture notes ai research assistant " scripture](#).

OpenAI. [Introducing chatgpt](#) [online]. 2022.

Eduard Petiska. 2023. Chatgpt cites the most-cited articles and journals, relying solely on google scholar’s citation counts. as a result, ai may amplify the matthew effect in environmental science. *arXiv preprint arXiv:2304.06794*.

Sundar Pichai. 2024. [Our next-generation model: Gemini 1.5](#).

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

$n \setminus i$	1	2	3	4	5	6	7	8	9	10
1	0.337	0.428	0.481	0.517	0.546	0.570	0.589	0.604	0.619	0.633
2	-	0.119	0.181	0.224	0.256	0.283	0.306	0.328	0.345	0.361
3	-	-	0.045	0.078	0.104	0.127	0.148	0.165	0.182	0.198
4	-	-	-	0.018	0.034	0.050	0.064	0.076	0.089	0.100
5	-	-	-	-	0.007	0.016	0.025	0.033	0.041	0.049
6	-	-	-	-	-	0.004	0.008	0.013	0.017	0.021
7	-	-	-	-	-	-	0.002	0.004	0.007	0.009
8	-	-	-	-	-	-	-	0.001	0.002	0.004
9	-	-	-	-	-	-	-	-	0.000	0.001
10	-	-	-	-	-	-	-	-	-	0.000

Table 6: This table shows the *coverage* as evaluated for various values of n and i on Bible data using rand_14730 and the cityblock distance metric, the random words topic model using the entire unstemmed vocabulary as topics; this was the most successful topic-based model.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Kim Renfro. 2020. 17 details you might have missed in the ‘harry potter’ books. *Business Insider*.

J.K. Rowling. 1997. *Harry Potter and the Philosopher’s Stone*. Bloomsbury Publishing, London, UK.

J.K. Rowling. 2003. *Harry Potter and the Order of the Phoenix*. Bloomsbury Publishing, London, UK.

Hamidah M Salleh. 2023. Errors of commission and omission in artificial intelligence: contextual biases and voids of chatgpt as a research assistant.

Brandon Sanderson. 2018. [What is your daily word-count/time goal?](#)

Mun’im Sirry. 2022. *The Qur’an with Cross-References*. De Gruyter.

Jerome H. Smith. 2016. *The Ultimate Cross-Reference Treasury, for e-Sword*. Thomas Nelson.

Leonid Taycher. [Books of the world, stand up and be counted! all 129,864,880 of you.](#) [online]. 2010.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. [Relic: Retrieving evidence for literary claims](#). *Preprint*, arXiv:2203.10053.

A Examples of Subtle Cross-references

Below we give two examples of cross-references more subtle than the Emperor’s New Groove callback in the Introduction.

A.1 Harry Potter Example

In J.K. Rowling’s first book *Harry Potter and the Philosopher’s Stone* (Rowling, 1997), at one point Harry thinks:

Could Snape possibly know they’d found out about the Philosopher’s Stone? Harry didn’t see how he could—yet he sometimes had the horrible feeling that Snape could read minds.

In the fifth Harry Potter book, *Harry Potter and the Order of the Phoenix* (Rowling, 2003), we find the following quote from Snape:

Those who have mastered Legilimency are able, under certain conditions, to delve into the minds of their victims and to interpret their findings correctly.

In other words, in an early book Harry wonders if Snape can read minds, and in a later book he finds out that Snape can read minds, at least after a fashion. However, unlike the callback from the

$n \setminus i$	1	2	3	4	5	6	7	8	9	10
1	0.341	0.436	0.494	0.534	0.564	0.589	0.611	0.628	0.642	0.655
2	-	0.114	0.176	0.221	0.258	0.289	0.313	0.336	0.354	0.372
3	-	-	0.039	0.073	0.099	0.122	0.143	0.162	0.179	0.195
4	-	-	-	0.013	0.027	0.042	0.056	0.069	0.081	0.093
5	-	-	-	-	0.005	0.012	0.019	0.027	0.033	0.041
6	-	-	-	-	-	0.002	0.005	0.009	0.012	0.016
7	-	-	-	-	-	-	0.001	0.002	0.004	0.006
8	-	-	-	-	-	-	-	0.000	0.001	0.002
9	-	-	-	-	-	-	-	-	0.000	0.000
10	-	-	-	-	-	-	-	-	-	0.000

Table 7: This table shows the *coverage* as evaluated for various values of n and i on Bible data using sentence embeddings and the cosine similarity metric, the random words topic model using the entire unstemmed vocabulary as topics.

Emperor’s New Groove given above, Harry’s supposition in the first book is not emphasized enough that it will easily be remembered 4 books later, nor does the fifth book refer back to the earlier thought, making this a much more difficult cross-reference to find.

We found this cross-reference in [Renfro \(2020\)](#), an article discussing this reference as well as other “foreshadowing” in the Harry Potter series.

A.2 Scriptural Example

One example of a more subtle cross-reference is given below between the Bible and the Book of Mormon—a book of scripture for The Church of Jesus Christ of Latter-day Saints.

John 9:1-3 – As he passed by, he saw a man blind from birth. And his disciples asked him, “Rabbi, who sinned, this man or his parents, that he was born blind?” Jesus answered, “It was not that this man sinned, or his parents, but that the works of God might be displayed in him.

After which, Jesus proceeds to heal the man of his blindness.

In Ether, God is speaking, and says:

Ether 12:27 – And if men come unto me I will show unto them their weakness. I give unto men weakness that they may be humble; and my grace is sufficient for all men that humble themselves before

me; for if they humble themselves before me, and have faith in me, then will make weak things become strong unto them.

The connections between the two passages are as follows:

- Both of these verses begin by mentioning a “flaw.” In John it is a physical disability, blindness, and in Ether it is the abstract concept of weakness.
- In each case the individual with the flaw is not blamed for the flaw. In John the idea of blame is specifically rejected by Jesus, and in Ether God assumes blame for weakness: “I give unto men weakness . . .”
- In each there is a different purpose given for the flaw. In John “that the works of God might be displayed,” and in Ether “that they may be humble.”
- Finally, the flaw is at the center of a transformation. In John Jesus heals the blind man so that he can see, and in Ether God promises, “I will make weak things become strong unto them.”

While there are very strong connections between these verses, they would not appear together in any kind of word-based search, and so would be difficult to identify without great familiarity with the texts.

	rand_20984 (city)	embeddings (cos)	ChatGPT (quote)
Author 1	0.36	0.70	0.16
Author 2	0.79	0.95	0.52

Table 8: Manually evaluated Jane Austen *coverage* results for rand_12753 (the entire unstemmed vocabulary), sentence embeddings, and ChatGPT given the source passage.

B Coverage Results for multiple values of n and i

See Table 6 for the *coverage* for various values of n and i for the unstemmed random word model with 14730 topics—the whole vocabulary—using the city-block distance metric, i.e. the best-performing topic-based model.

See Table 7 for the *coverage* for various values of n and i for the sentence embedding model using cosine similarity.

Remember that i is the number of suggested cross-references evaluated, and n is the minimum number of valid cross-references that must be found for a source passage of text to be considered covered. In other words $n = 2$ and $i = 2$ means that two cross-references are suggested and both must be valid cross-references in order for that passage to add to the overall *coverage*. The bottom left side of the table is blank because it is impossible to find more valid cross-references than verses examined, so n can never be larger than i and produce a valid result.

It may be of interest to note that although the sentence embedding model performs better than the rand_14730 model at $n = 1$, $i = 5$, as the value of n increases, rand_14730 performs better than the sentence embedding model including for some values of i when the value of n is 2. By the time n is 3, rand_14730 outperforms the sentence embedding model for all values of i .

C Finetuning ChatGPT

We used the following prompt to finetune ChatGPT with the text of *Sons of the Starfarers*:

What does the [NTH] chapter, [CHAPTER-NAME] of [TITLE] by Joe Vasicek say? It’s important that you know this so you can cross-reference it later. [CHAPTERTEXT]

We replaced [NTH] with the ordinal number of the chapter, [CHAPTERNAME] with the name of the chapter, [TITLE] with the title of the book, and [CHAPTERTEXT] with the actual text of the

chapter. Note that when using a similar prompt with *Pride and Prejudice*, ChatGPT is capable of reproducing Jane Austen’s writing.

We trained on a total of 1.3 million tokens, for 3 epochs with a batch size of 1, and a learning rate multiplier of 2—these were the default settings suggested by OpenAI. The total cost to finetune was \$10.49 USD.

D Jane Austen Hand-evaluation

See Table 8 for the numerical results of our authors hand-evaluating the Jane Austen cross-reference suggestions.

Extracting position titles from unstructured historical job advertisements

Klara Venglarova
University of Graz

Raven Adam
University of Graz

Georg Vogeler
University of Graz

klara.venglarova@uni-graz.at raven.adam@uni-graz.at georg.vogeler@uni-graz.at

Abstract

This paper explores the automated extraction of job titles from unstructured historical job advertisements, using a corpus of digitized German-language newspapers from 1850-1950. The study addresses the challenges of working with unstructured, OCR-processed historical data, contrasting with contemporary approaches that often use structured, digitally-born datasets. We compare four extraction methods: a dictionary-based approach, a rule-based approach, a named entity recognition (NER) mode, and a text-generation method. The NER approach, trained on manually annotated data, achieved the highest F1 score (0.944 using transformers model trained on GPU, 0.884 model trained on CPU), demonstrating its flexibility and ability to correctly identify job titles. The text-generation approach performs similarly (0.920). However, the rule-based (0.69) and dictionary-based (0.632) methods reach relatively high F1 Scores as well, while offering the advantage of not requiring extensive labeling of training data. The results highlight the complexities of extracting meaningful job titles from historical texts, with implications for further research into labor market trends and occupational history.

1 Introduction

Historical job advertisements provide unique information about the history and development of the labor market. Analyzing the positions offered and sought over time offers insights into temporal and regional differences and development, as well as into social aspects, such as gender-specific job offers. The first step in such an analysis is the extraction of job titles. However, when using historical data from digitized newspapers, rather than digitally-born structured data, the automatic extraction of job titles proves to be a non-trivial task.

In the JobAds Project (FWF P35783), we study

historical job advertisements from digitized newspapers from the ANNO corpus (Österreichische Nationalbibliothek, 2021). The advertisements are predominantly in German, and our defined time span is 1850-1950. The newspaper pages were initially obtained in the form of images and transformed into textual data by conducting the processes of page segmentation, optical character recognition (OCR), and automatic post-correction based on manually transcribed ground truth. Afterwards, the job advertisements were extracted. A result of such a pipeline is a corpus containing tens of thousands machine-readable yet unstructured job advertisements.

Contemporary research often works with modern, digitally-born data, and usually benefits from their structure, such as HTML tags, to identify position titles. Modern research in the context of job advertisements addresses challenges such as extracting or grouping requirements in the job ads, e.g. (Gnehm et al., 2022; Ternikov, 2022; Grüger & Dr. Schneider, 2019; Wowczko, 2015; Litecky et al., 2010), automated matching process between the position requirements and the skills of a candidate written in their CV, e.g. (Fernández-Reyes & Shinde, 2019; Sayfullina et al., 2018; S. Chala et al., 2017; Guo et al., 2016; J. Malinowski et al., 2006), job advertisements classification/categorization, e.g. (Gnehm & Clematide, 2020; Boselli et al., 2018; Malherbe et al., 2015; Amato et al., 2015), and job title classification, e.g. (Colace et al., 2019; Boselli et al., 2017; Zhu et al., 2017); with some works covering more than one research focus. In contrast, we face the challenge of working with unmarked text, and need to find an automated way to extract this information.

In this paper, we present a comparison of four approaches to identification of position titles in historical job advertisements. On the one hand, we consider dictionary-based and rule-based approaches, which do not require a time-consuming creation of

an annotated training dataset. On the other hand, we use a machine-learning (ML) approach, specifically named entity recognition (NER) and text-generation models training. This requires manually annotated training data, but yields better results and can recognize position titles beyond those seen in the training data. Please note that in this paper, we aim to extract position titles appearing inside the advertisement, in contrast to some of the aforementioned research works which aimed to classify jobs into the occupation categories.

The following section summarizes existing research related to job title identification and information retrieval from unstructured text. Section 3 describes our dataset and the process of job titles annotation. Section 4 presents in detail various approaches we used to identify the job titles and the evaluation methodology, while section 5 presents and discusses the results. Section 6 concludes this paper.

2 Related Work

While most of the modern research benefits from the structure of ads to extract job titles, as exemplified above, some research addresses the challenge of their extraction nevertheless. One reason for this can be the noisy information included within the job title tags, such as the name of the company, or the need to work with ads from newspapers that were obtained in the form of the image.

Rahhal et al. (2023) develop a methodology for matching a job ad with its standardized occupation in French language. Although the job title is included within a dedicated HTML tag, the authors further process it because other words, such as ‘looking for’, are sometimes included. They remove these extraneous words based on a manually created to-delete list to obtain cleaned job titles.

Bandara et al. (2021) work with unstructured text, as they scrape job ads in the form of images from job web portals and newspapers. They apply an OCR process to convert them into machine-readable text. Aiming to create a structured dataset, they extract information such as the position name, skills, company name, and contact information. They use a rule-based approach, matching regex patterns or phrases. However, the accuracy of job title identification is only 56% (Bandara et al., 2021, p. 148). The authors do not explicitly state which specific regex patterns were used for job titles extraction.

Neculoiu et al. (2016) focus on job title normalization, where job titles are normalized according to a predefined set of occupations. Researchers dealing with this question also need to extract job titles; however, they often rely on external taxonomies, structured text, or manual labeling. In (Neculoiu et al., 2016, p. 152), “the job titles were manually and semiautomatically collected from resumes and vacancy postings.”

Not having found a suitable approach in the existing literature, we focus on other ways to extract job titles from the advertisements. Information extraction (IE) is the process of automatically extracting entities, objects, and their roles from text, often within a specific domain or topic (Hobbs & Riloff, 2010), with named entity recognition (NER) being one of the common technologies used (Tjong Kim Sang & De Meulder, 2003; Collins & Singer, 2002; Cucerzan & Yarowsky, 2001). While the standard entities in NER models are typically proper names, locations, or dates, custom NER models can be trained to recognize new entities. Therefore, we can consider job titles entities, and having created appropriate training data, use the same approach to extract job titles from the unstructured text.

The emergence of large language models (LLMs) has also opened up the possibility to approach NER as a text generation or translation task (Keraghel et al., 2024). While this approach has, to the best of our knowledge, not been evaluated on historical data, it has seen promising results in a variety of fields, e.g., (Tavan & Najafi, 2022; Wang et al., 2023), and we shall include this method in our evaluation.

While no work specifically addresses position extraction from historical data, several works deal with historical data and NER. Grover et al. (2008) addressed recognition of person and place names within a digitized corpus of British parliamentary proceedings from 1685–1691 and 1814–1817 using a rule-based approach. Working with the output of an OCR software, they reach total f-scores from 70.35 to 76.94 on individual datasets (Grover et al., 2008, p. 1346). The main challenge were the OCR mistakes, namely the noise, misrecognition of characters and issues with separating the text from the marginal notes.

Won et al. (2018) focus on toponyms in two collections of historical letters, one collection in early-modern English, another in modern English. As obstacles they mention e.g. “language changes over time, spelling variations, OCR errors, sources

written in multiple languages, and general ambiguity in language use” (Won et al., 2018, p. 2). The authors do not train new models but evaluate the performance of existing modern tools for the task of location extraction, including two different types of pre-processing and experimenting with re-writing early-modern English text into modern English. They evaluate five different NER systems and an ensemble method which works based on a voting system among the individual NER models. All experiments reached a best minimum F1 score of about 70 (Won et al., 2018, p. 8).

Labusch et al. (2019) trained a pre-trained BERT model for a NER task in contemporary and historical German corpora, containing entities of location, organisation and person. Working with historical corpora is hindered through the less standardized languages and errors in the OCR. Combining unsupervised pre-training on historical German corpus with supervised pre-training using contemporary German NER ground-truth, they achieve the highest F1 score of 84.6 on historical data for 5-fold cross validation.

3 Dataset

Using digitized historical newspapers from the ANNO corpus, we manually annotated, OCRed and corrected several thousands job advertisements spanning the period 1850-1950 from 14 different newspapers. Within a subset of these job ads, we annotated position names using the doccano software (Nakayama et al., 2018), yielding 1,486 job advertisements as training data and 637 as testing data. The split into training and testing dataset was random.

Ads that did not explicitly mention a position or contained errors, such as inconsistent spacing, were mostly not included within the datasets for the evaluation task (Tab. 1), however, after the training and evaluation during a human control we discovered a small number of ads containing errors that were included in the dataset by mistake. Additionally, we standardized the text by replacing the *long s* with the letter *s* in all ads. The training data was used to train the NER and text-generation models and also served as a basis for creating part-of-speech (POS) and syntax rules, as well as for adding entries to the dictionary, both established through human observations. The testing data was consistently used for evaluation purposes.

Although all duplicate ads were removed and the

training and testing datasets are mutually exclusive, we cannot exclude the possibility that very similar ads appear in both datasets, given the homogeneity of ads and their structure defined by their genre. Each advertisement could contain more than one position offered/sought for.

4 Methods

Based on preliminary results, four approaches were selected and compared: a dictionary-based approach, a rule-based approach, a NER and a text generation approach.

4.1 Dictionary-based Approach

The dictionary-based approach searches for position titles based on the database of historical occupations HISCO¹ (Leeuwen et al., 2002). The HISCO (historical international classification of occupations) database is based on the coding of 1,000 most frequent male and female occupational titles in datasets from Belgium, Britain, Canada, France, Germany, the Netherlands, Norway and Sweden. It contains titles included in parish and civil registration documents (International Institute of Social History, 2023). The dictionary of German occupations contains 1297 job titles and was subsequently further enriched based on our observations by collocations (e.g. *Mädchen für alles* [girl servant, lit. girl for everything], *Stütze der Hausfrau* [housewife’s help]) and positions related to apprenticeship (e.g. *Lehrmädchen* [apprentice f.], *Praktikant* [intern]), which we also aim to identify. Every exact match of the collocations was identified as a position.

To overcome difficulties of different spelling variants (*Commis* and *Kommis* [assistance]), we lemmatized every entry in the dictionary as well as each token in the advertisement using the DTA::CAB web-service (Jurish, 2012)² which serves for an ‘error-tolerant linguistic analysis for historical German text’. If a lemmatized token matched a position from the dictionary, regardless of the upper/lowercase, the entire word containing this string was identified as a position. Even when this approach introduces some false positives (e.g., identifying *Architektur* [architecture] because it contains *Architekt* [architect]), it helps to identify true positives that are not in the dictionary (e.g., *Steinbrechermeister* [stone crusher mas-

¹<https://iisg.amsterdam/en/data/data-websites/history-of-work> [15.6.2024]

²<https://www.deutschestextarchiv.de/demo/cab/> [26.8.2024]

Advertisement text	Part of the corpus	Annotated entities
Kinderliebendes Mädchen zu 2 Kindern und Mithilfe dringend gesucht. Zuschriften unter „Kinderliebend 2148“ an Rasteiger, Joanneumring 6. 1559 [Child-loving girl urgently wanted to 2 children and assistance. Send letters under ‘Kinderliebend 2148’ to Rasteiger, Joanneumring 6. 1559]	No (no position title)	-
Suche eine tüchtige Wirt schafterin . 35—40 J. spätere Ehe nicht ausgeschlossen. Un ter „Ehrlich 2270“ Neue Zeit [Looking for a hard-working housekeeper . 35-40 yrs. later marriage not excluded. Under Sincere 2270’ Neue Zeit]	No (wrong spacing)	-
Ein Commis in einer Eienhandlung wird acceptirt. 2936 [An assistance will be accepted in an ironmonger. 2936]	Yes	Commis
Maschinenschlosser resp. Automatenschlosser wird gesucht. Vorzustellen Ing. Gasser-Steiner Graz Strauchergasse 16. [Machine locksmith or automatic locksmith is wanted. To be introduced at Ing. Gasser-Steiner Graz Strauchergasse 16.]	Yes	Maschinenschlosser, Automaten-schlosser

Table 1: Example of annotated position names in job advertisement and of advertisements excluded from our dataset.

ter], *Tapezierergehilfen* [paperhanger’s assistant], *Weißnäherin* [seamstress]), often because they are composed of more words that specify the position. The number of positions identified is equal to the number of matches.

4.2 Rule-based Approach

The rule-based approach benefits from the strong linguistic structure of job ads. First, the same collocations are searched for as in the dictionary-based approach (e.g., *Mädchen für alles* [girl servant, lit. girl for everything], *Stütze der Hausfrau* [housewife’s help]). If a match is found, the collocations are identified as positions. If no match is found, the word ‘*als*’ [as] (or ‘*Als*’ [As]) is searched with whitespaces as stated. If this word is found, the first noun after it is identified as a position, e.g.:

Suche Stelle als Hausgehilfin in Bäckerei od. Gastwirtschaft. [Looking for a job as a maid in a bakery or restaurant.]

If this search is unsuccessful, the search continues for the word ‘*stelle*’ [position]. Note that in German, *stelle* beginning with lowercase can only be found in the middle of a word. If this is found, the string containing it is considered a position, e.g.:

Alleinstehende Frau mit kl. einjähr. Buberl bittet

um Hausmeisterstelle. [Single woman with a small one-year-old boy asks for a caretaker **position**.]

If none of these conditions is met, the ad is searched for the first noun which is subject, root, or conjunct in the sentence. These syntactic roles were chosen based on observations in training data, including the observed miss-classifications caused probably by concise and archaic language. This noun is predicted as a position. If it is followed by ‘*und*’ [and] or ‘*oder*’ [or] and the word after this conjunction is a noun, this second noun is considered another position.

The advantage of this approach is its independence from an external dictionary, spelling variations and words that do not appear in the dictionary. The main drawback is handling advertisements that offer more than one position, as it is hard to define rules that include multiple true positives without introducing false positives. In the testing data, 514 instances contain one single position, while 123 of them contain more than one position. For the POS tagging and dependency parsing, the SpaCy library (Honnibal & Montani, 2017) with the ‘*de_core_news_lg*’ model was used.

4.3 Named Entity Recognition Approach

The NER approach treats the positions in the advertisements as named entities and attempts to identify them accordingly. As positions are not among

the standard entities, a custom model needs to be trained.

We trained two different models, one with the Token2Vec architecture, and one with a transformer-based architecture. The `spacy.Tok2Vec.v2` model contains two steps: it creates context-independent word vector representation, and it encodes context into the embeddings, using architecture like a CNN, BiLSTM or transformer (SpaCy, n.d.). In comparison, the `spacy-transformers.TransformerModel.v3` uses transformer models from the HuggingFace transformers library to create more advanced, context-aware embeddings, leading to better performance in tasks like NER (SpaCy, n.d.). Training on GPU allows for faster training times, particularly for transformer-based models, which involve complex matrix operations (Kelleher, 2019, pp. 92–94), and they reach generally higher performance than training on CPU.

For the training, we used default SpaCy settings. In both cases, the language is set to ‘German’, and the optimization is set to efficiency. In the first case, we set the hardware to CPU, while in the second case, we select GPU. For further details on the training parameters, we point readers to SpaCy official documentation³.

In both cases, the NER model was trained on the training dataset and evaluated on the testing dataset, as specified in the Dataset section. The advantage of this approach is some ability to generalize; however, the disadvantage is the need for manual training data creation, which can be time-consuming, and may limit model’s effectiveness if the training data is not selected appropriately.

4.4 Text Generation Approach

Whereas the NER approach identifies and extracts entities within text, the text generation approach creates new text that directly facilitates the identification of relevant entities. In the chosen approach a given text, e.g. “Machine locksmith or automatic locksmith is wanted.” is rewritten into “<Machine locksmith> <automatic locksmith>”. If no position is found within a given text, it is rewritten into “none”.

We used a `hmByT5` model⁴ as our base model due to two major reasons. The first one being that the `hmByT5` models are all trained on multilingual historical data, which overlaps with the time

period of our dataset. The second one being that `ByT5` encoded text byte-wise instead of the word or subword level. Therefore, the model requires no internal vocabulary and is more adaptable to words that were not included in the base model training.

Because a significant number of job ads exceeds the maximum encoding length of the chosen model, all ads concerned were split into segments of up to 120 bytes for training and evaluation. Afterwards the model was trained for 15 epochs and the epoch with the highest `sacreBLEU` score was chosen as the final model. Due to splitting up ads into smaller segments, the model is evaluated in two different ways (Tab. 2). First, based on the prediction for each segment. Since not all segments contain positions, properly predicting “none” as a result is included when calculating evaluation metrics. The second evaluation re-merges the predictions from all segments into the original ads and evaluates only the predicted positions without including “none” predictions, as this is most comparable to the NER approach.

4.5 Evaluation

The evaluation involves a pre-processing step, in which we aim to standardize both the annotated positions, and the predicted ones, to avoid penalizing a model for e.g., including white spaces. First, we delete the word ‘stelle’ [*position*] if it is included, e.g. *Hausmeisterstelle* [*caretaker position*] becomes *Hausmeister* [*caretaker*]. Afterwards, both annotated and predicted positions are lemmatized using the `DTA::CAB` web-service (for details, see section 4.1).

In the next step, the two lists of annotated and predicted positions are compared for each advertisement. Certain tolerance is included by allowing Levenshtein distance (Levenshtein, 1965) of 0.1 between the two strings in order not to penalize the generative approach if it makes minor mistakes, such as generating (position> instead of <position>. The number of True Positives, False Positives and False Negatives is identified as follows:

- **True Positives (TP):** The model correctly predicts a position that is present in the list of annotated positions.
- **False Positives (FP):** The model predicts a position that is not present in the list of annotated positions.
- **False Negatives (FN):** The model misses a

³<https://spacy.io/usage/training> [29.7.2024]

⁴<https://huggingface.co/hmbyt5-preliminary/byt5-small-historic-multilingual-span20-flax> [25.8.2024]

Advertisement segment	Model output	Aggregation of entire output
Aelterer Herr, Beamter i. P. gesund und rüstig, alleinstehend sucht Posten in Schloß, Kloster [Older gentleman, civil servant , healthy and sprightly, single seeks position in castle, monastery]	<Beamter>	<Beamter>, <Ausseher>, <Pfortner>, <Hausgehilfe>
als Ausseher , Pfortner , Hausgehilfe zu Gartenarbeiten auch auswärts. Gute Zuschriften [as external , doorman , housekeeper for gardening work also away from home. Good applications]	<Ausseher>, <Pfortner>, <Hausgehilfe>	<Beamter>, <Ausseher>, <Pfortner>, <Hausgehilfe>
unter: "Vertrauenswürdig und verlässlich Nr 85368" an das Welt-Blatt. Wien, I. Schulerstraße. [under: "Trustworthy and reliable no. 85368W to the Welt-Blatt. Vienna, I. Schulerstrasse.]	none	<Beamter>, <Ausseher>, <Pfortner>, <Hausgehilfe>

Table 2: Example of segmented job advertisement along with text generation model output and aggregation of position names found for the whole add without including "none" predictions. Evaluation is performed directly on the outputs as well as the aggregated predictions.

position that is present in the list of annotated positions.

Using this information, we calculate F1 Score, Recall, and Precision (Powers, 2011), where the metrics are calculated as follows:

- **F1 Score:** $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$
- **Recall:** $\text{TP} / (\text{TP} + \text{FN})$
- **Precision:** $\text{TP} / (\text{TP} + \text{FP})$

This evaluation is identical for all four approaches, making their scores comparable.

5 Results and Discussion

Table 3 presents the results using F1 Score, Recall, and Precision for the above described methods on the testing dataset.

The NER approach (Fig. 1) reached the highest F1 score of 0.944 using the transformer architecture, resp. 0.884 when trained on CPU on the testing dataset, indicating that this method is more flexible than the rule-based and dictionary-based. However, a significant drawback of this approach is its reliance on a manually created training set.

The generative approach shows comparable performance with an F1 score of 0.920 when evaluated based on segments and 0.902 when evaluated on the aggregated results for entire ads. Since the only difference between the two evaluation approaches is the exclusion of "none" predictions when aggregating, the decrease from the segment based score

Fesche solide 12599 Kassierkellnerin **position** und tüchtige Köchin **position** finden sofort
Stellung. Stadtparkrestaurant, Saaz.

Figure 1: Example of positions identified by the NER model. [Good-looking, solid 12599 cashier waitress and hard-working cook are wanted immediately. City park restaurant, Saaz.]

to the ad based score indicates that the model is slightly better suited to determine that no positions are mentioned in a line of text instead of extracting all mentioned positions. A preliminary check of prediction mistakes has, however, shown that some false positives occur from identifying ambiguous words such as *Mann* [man], *Mädchen* [girl] or *Französinen* [French women], which are also used to advertise positions but were not labeled in our dataset as well as words with misplaced spacing, such as *Wirt schafterin* [housekeeper], which was also not included as a labeled position. This behavior indicates potentially interesting and useful generalization ability.

The rule-based approach appears to be well-suited for this task, given the strong linguistic structure of job advertisements. It is also immune to certain spelling variations or typographic mistakes, as long as they do not include spacing errors, or do not hinder the correct POS classification. The problematic aspect of this approach is when more than one position is advertised within an ad, leading to either incomplete detection of all positions or the introduction of false positives in other ads. Another technical obstacle is the misclassification

Method	F1 score	Recall	Precision
Dictionary-based approach	0.632	0.646	0.617
Rule-based approach	0.690	0.613	0.789
NER approach (CPU)	0.884	0.866	0.903
NER approach (GPU)	0.944	0.932	0.956
Text Generation approach (segment)	0.920	0.918	0.922
Text Generation approach (whole ad)	0.902	0.894	0.909

Table 3: Results for different approaches for predicting position names on the testing dataset.

of parts-of-speech or dependencies due to archaic and elliptic language used within historical job advertisements, while the modern tools are generally designed for modern data and contemporary language. They may also not handle advertisements containing a large number of abbreviations, which is often the case because every line in a newspaper was costly.

The dictionary-based approach reached the lowest F1 score. While lemmatizing/standardizing of the tokens mainly solves the issue of spelling variants, it is time-consuming to lemmatize the entire text of the advertisement. However, this approach without lemmatization reaches lower score as usually only one form is present in the dictionary, while the text contains several variants (e.g., *Commis* and *Kommis* [assistance], *Kontoristin* and *Comptoiristin* [clerk f.]). Additionally, some professions are simply not present in the dictionary, especially those that were less common or highly specialized. This approach also fails when typographic errors are present in the text (*Kö chin* with line-breaker in the original text instead of *Köchin* [cook f.]). Moreover, all mentioned positions are identified, even if they appear as part of a name of a street, or if they are not the primary focus of the advertisement but just mentioned within it, e.g.:

*Kinderarzthilfe. mit zweijähriger Praxis, sucht Stelle bei Arzt oder in einem Laboratorium. [Pediatric assistant with two years of experience is looking for a position with a **doctor** or in a laboratory.]*

Our dataset contained advertisements in their ideal form, i.e., manually corrected with only occasional typographic errors. In reality, the thousands of ads can only be post-corrected automatically rather than manually, and more errors will thus be present in the data. This will mostly affect the dictionary-based approach, which needs exact matches with correct spelling, and partially the rule-

based approach, which is in theory immune to the spelling variations but fails to correctly identify the POS and syntax dependencies if too many errors occur.

The last consideration is the ambiguous nature of what a job title in historical job advertisements is. Many ads contain words like *Mann* [man], *Mädchen* [girl] without specifying a clear job title, although it is implicitly understood that e.g., the girl is wanted to help in the household. Another example is *Französinnen* [French women], mentioned alongside other job titles for women, where it is implied that they are sought to teach the French language or to provide companionship while conversing in French (Fig. 2). These advertisements were not included in our dataset for this evaluation task, however, they appear commonly in the corpus.

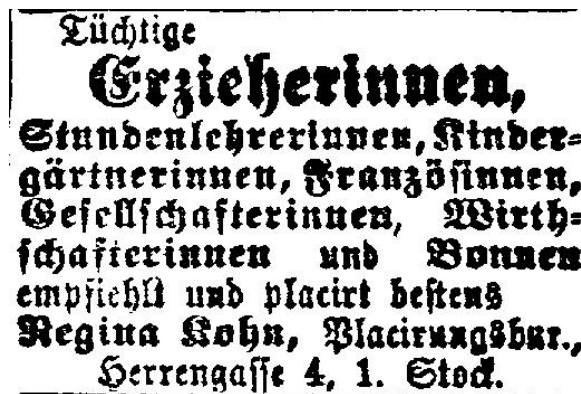


Figure 2: [Regina Kohn, Placement Office, Herrengasse 4, 1st floor; recommends and places competent educators, class teachers, kindergarten teachers, French women, companions, housekeepers, and maids.] Source: Prager Tagblatt, 23.8.1896, p. 30, <https://anno.onb.ac.at/cgi-content/anno?aid=ptb&datum=18960823&seite=30>

Further research could help to address these limitations by focusing on the semantic context of the advertisements, perhaps using techniques like average embeddings to represent the average meaning

of the whole advertisement. This approach could help overcome the challenges posed by the ambiguous and context-dependent nature of job titles in historical advertisements. Also, the potential of text-generation methods shall be further investigated.

6 Conclusion

Historical job advertisements offer many opportunities to study the transformation of the labor market. The first step in analyzing these ads is the extraction of job titles, which allows for the exploration of positions offered and sought, their frequency analysis, and temporal and regional variations. In this study, we compared four approaches for job title extraction: a dictionary-based approach, a rule-based approach, a NER approach, and a text-generation approach. The NER approach achieved the highest F1 score of 0.944 when a transformer-based model was trained on GPU, resp. 0.884 when a model was trained on CPU, as we treated position names as entities and trained a model specifically for their identification. The generative model follows closely with a F1 score of 0.920 and offers potential viability for identifying ambiguous positions. The rule-based (0.69) and dictionary-based (0.632) approaches had lower F1 Scores but offered the advantage of not requiring the time-consuming creation of a training dataset.

While extracting position names is a crucial first step, two important considerations remain:

1. Not all job advertisements explicitly state the position being offered or sought.
2. As noted in (Wowczko, 2015, p. 36), “job titles can oftentimes be ambiguous and fail to reflect the true nature of the work”. This is particularly true for generic titles such as *Arbeiterin* [worker f.], *Bedienerin* [servant f.], *Praktikant* [intern], *Lehrling* [apprentice], which do not express sufficient information about the work involved.

Given these challenges, our future research will focus on grouping job advertisements by sector and exploring methods to predict job titles based on the job description, as demonstrated by (Huynh et al., 2019). This approach could help address the ambiguity and lack of specificity often found in historical job titles.

Limitations

One limitation encountered during our work was a lack of definition of what a job title is. While in most ads, the titles are explicit, there is a not-negligible number of ads containing just words like *Mädchen* [girl] or *Mann* [man]. The lack of definition caused minor inconsistencies among annotators which may slightly skew the results. It is also important to note, that the two machine learning approaches, namely NER and text generation, are not trained on the same task. The NER approach is trained on token classification and predicts the exact location of a job position. Whereas the text generation approach rewrites the original text to only consist of job positions. While this difference is not of major significance for our specific task, this can be relevant to consider for potential use cases. Another limitation is found during the lemmatization step. Only a very limited number of resources for lemmatization of historical german texts are currently available and are mostly rule based. As such, job positions may not be properly normalized for evaluation, when they are not covered by the utilized set of rules.

Acknowledgments

We thank the Austrian National Library (ÖNB) for providing the data, Saranya Balasubramanian for a great help with its processing, and Wiltrud Mölzer, Jörn Kleinert, Meike Linnewedel, Clara Hochreiter and Melanie Frauendorfer for their efforts in correcting and annotating it. We are also thankful to Vladimir Matlach for his consultations and helpful insights. This work was supported by the FWF under grant number P 35783.

Code and Data Availability

The code and data containing advertisements text and annotated positions, are available at <https://github.com/JobAds-FWFProject/PositionsExtraction>.

References

- Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., & Picariello, A. (2015). Classification of Web Job Advertisements: A Case Study. 144–151.
- Bandara, R., Gunasekara, H., Peiris, W., Wijekoon, W., De Silva, T., Hewawalpita, S., & Rathnayake,

- H. (2021). Information extraction from Sri Lankan job advertisements via rule-based approach.
- Boselli, R., Cesarini, M., Mercurio, F., & Mezzanzanica, M. (2017). Using machine learning for labour market intelligence. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part III* 10, 330–342.
- Boselli, R., Cesarini, M., Mercurio, F., & Mezzanzanica, M. (2018). Classifying online Job Advertisements through Machine Learning. *Future Generation Computer Systems*, 86, 319–328. <https://doi.org/10.1016/j.future.2018.03.035>
- Colace, F., De Santo, M., Lombardi, M., Mercurio, F., Mezzanzanica, M., & Pascale, F. (2019). Towards labour market intelligence through topic modelling.
- Collins, M., & Singer, Y. (2002). Unsupervised Models for Named Entity Classification. *Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC*.
- Cucerzan, S., & Yarowsky, D. (2001). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.
- Fernández-Reyes, F. C., & Shinde, S. (2019). CV Retrieval System based on job description matching using hybrid word embeddings. *Computer Speech & Language*, 56, 73–79. <https://doi.org/10.1016/j.csl.2019.01.003>
- Gnehm, A.-S., Bühlmann, E., Buchs, H., & Clematide, S. (2022). Fine-Grained Extraction and Classification of Skill Requirements in German-Speaking Job Ads. *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, 14–24. <https://doi.org/10.5167/uzh-230653>
- Gnehm, A.-S., & Clematide, S. (2020). Text Zoning and Classification for Job Advertisements in German, French and English. In D. Bamman, D. Hovy, D. Jurgens, B. O’Connor, & S. Volkova (Eds.), *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science* (pp. 83–93). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.nlpccs-1.10>
- Grüger, J., & Dr. Schneider, G. (2019). Automated Analysis of Job Requirements for Computer Scientists in Online Job Advertisements (p. 233). <https://doi.org/10.5220/0008068202260233>
- Guo, S., Alamudun, F., & Hammond, T. (2016). Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, 60, 169–182. <https://doi.org/10.1016/j.eswa.2016.04.013>
- Hobbs, J. R., & Riloff, E. (2010). Information Extraction. *Handbook of Natural Language Processing*, 2.
- Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
- Huynh, T., Nguyen, K., Nguyen, N., & Nguyen, A. (2019). Job Prediction: From Deep Neural Network Models to Applications.
- I. Rahhal, K. M. Carley, I. Kassou, & M. Ghogho. (2023). Two Stage Job Title Identification System for Online Job Advertisements. *IEEE Access*, 11, 19073–19092. <https://doi.org/10.1109/ACCESS.2023.3247866>
- Grover, C., Givon, S., Tobin, R., & Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 1343–1346.
- International Institute of Social History. (2023). *History of Work—HISCO*. International Institute of Social History. <https://iisg.amsterdam/en/data/data-websites/history-of-work>
- Kelleher, J. D. (2019). *Deep learning*. MIT press.
- Labusch, K., Zu, S., Kulturbesitz, B., Neudecker, C., & Zellhöfer, D. (2019, October). *BERT for Named Entity Recognition in Contemporary and Historical German*.
- J. Malinowski, T. Keim, O. Wendt, & T. Weitzel. (2006). Matching People and Jobs: A Bilateral Recommendation Approach. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS’06)*, 6, 137c–137c. <https://doi.org/10.1109/HICSS.2006.266>
- Jurish, B. (2012). Finite-state Canonicalization Techniques for Historical German. [Universität Potsdam]. urn:nbn:de:kobv:517-opus-55789
- Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. <https://arxiv.org/abs/2401.10825>

- Leeuwen, M. H. D. van, Edvisson, S., Maas, I., & Miles, A. (2002). HISCO: Historical International Standard Classification of Occupations. <https://iisg.amsterdam/en/data/data-websites/history-of-work>
- Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics. Doklady*, 10, 707–710.
- Litecky, C., Aken, A., Ahmad, A., & Nelson, H. (2010). Mining for Computing Jobs. *Software, IEEE*, 27, 78–85. <https://doi.org/10.1109/MS.2009.150>
- Malherbe, E., Cataldi, M., & Ballatore, A. (2015). Bringing Order to the Job Market: Efficient Job Offer Categorization in E-Recruitment.
- Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text Annotation Tool for Human. <https://github.com/doccano/doccano>
- Neculoiu, P., Versteegh, M., & Rotaru, M. (2016). Learning Text Similarity with Siamese Recurrent Networks. <https://doi.org/10.18653/v1/W16-1617>
- Österreichische Nationalbibliothek. (2021). ANNO Historische Zeitungen und Zeitschriften. <https://anno.onb.ac.at/>
- Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.
- S. Chala, S. Harrison, & M. Fathi. (2017). Knowledge extraction from online vacancies for effective job matching. 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 1–4. <https://doi.org/10.1109/CCECE.2017.7946793>
- Sayfullina, L., Malmi, E., & Kannala, J. (2018). Learning Representations for Soft Skill Matching. In W. M. P. van der Aalst, V. Batagelj, G. Glavaš, D. I. Ignatov, M. Khachay, S. O. Kuznetsov, O. Koltsova, I. A. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. M. Pardalos, M. Pelillo, & A. V. Savchenko (Eds.), *Analysis of Images, Social Networks and Texts* (pp. 141–152). Springer International Publishing.
- SpaCy. (n.d.). *Model Architectures*. <https://spacy.io/api/architectures#Tok2Vec> [3.10.2024]
- SpaCy. (n.d.). *Model Architectures*. <https://spacy.io/api/architectures#TransformerModel> [3.10.2024]
- Tavan, E., & Najafi, M. (2022). MarSan at SemEval-2022 Task 11: Multilingual complex named entity recognition using T5 and transformer encoder. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, & S. Ratan (Eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* (pp. 1639–1647). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.semeval-1.226>
- Ternikov, A. (2022). Soft and hard skills identification: Insights from IT job advertisements in the CIS region. *PeerJ Computer Science*, 8, e946.
- Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147. <https://aclanthology.org/W03-0419>
- Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. <https://arxiv.org/abs/2304.10428>
- Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5. <https://doi.org/10.3389/fdigh.2018.00002>
- Wowczko, I. (2015). Skills and Vacancy Analysis with Data Mining Techniques. *Informatics*, 2, 31–49. <https://doi.org/10.3390/informatics2040031>
- Zhu, Y., Javed, F., & Ozturk, O. (2017). Document embedding strategies for job title classification. *The Thirtieth International Flairs Conference*.

Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs

Natalie Hervieux^{1*} Peiran Yao^{1*} Susan Brown² Denilson Barbosa¹

¹University of Alberta ²University of Guelph

{nhervieu,denilson}@ualberta.ca

Abstract

The digital humanities (DH) community fundamentally embraces the use of computerized tools for the study and creation of knowledge related to language, history, culture, and human values, in which natural language plays a prominent role. Many successful DH tools rely heavily on Natural Language Processing methods, and several efforts exist within the DH community to promote the use of newer and better tools. Nevertheless, most NLP research is driven by web corpora that are noticeably different from texts commonly found in DH artifacts, which tend to use richer language and refer to rarer entities. Thus, the near-human performance achieved by state-of-the-art NLP tools on web texts might not be achievable on DH texts. We introduce a dataset¹ carefully created by computer scientists and digital humanists intended to serve as a reference point for the development and evaluation of NLP tools. The dataset is a subset of a born-digital textbase resulting from a prominent and ongoing experiment in digital literary history, containing thousands of multi-sentence excerpts that are suited for information extraction tasks. We fully describe the dataset and show that its language is demonstrably different than the corpora normally used in training language resources in the NLP community.

1 Introduction

The digital humanities (DH) research community makes up a large user base for natural language processing (NLP) tools and algorithms (McGillivray et al., 2020; Biemann et al., 2014). Digital humanists have long been using cultural heritage data for meaningful NLP work, where NLP in DH includes everything from linguistic analysis of change over time within large linguistic corpora (Schlechtweg et al., 2020) to narratology (Piper et al., 2021) to

literary history (Underwood et al., 2018) and stylometry (Stamatatos, 2009).

However, there are risks associated with LLMs that are particularly relevant to DH. Unlike the average web document, texts in the humanities tend to use rich and complex writing styles, historical language, and references to under-represented long-tail entities (Olieman et al., 2017; Nurmikko-Fuller, 2023). LLMs have known problems with bias towards the contemporary and popularity bias (Dai et al., 2024). Chen et al. (2024) warn of a "Spiral of Silence" where over time, by iteratively training on LLM-generated content, LLM-based retrieval systems deprioritize accurate human-generated content and lose diversity in the information they return. If future NLP is dominated by LLMs that ignore the outliers that are so important to humanities scholarship (D'Ignazio, 2021; Jockers, 2013), this will negatively impact humanities research, our sense of history, and the public. As Brown and Simpson (2013) assert, "marginality and uniqueness are what humanities scholars often seek to discover and analyse". We need curated datasets for evaluating and fine-tuning LLMs with the priorities and expertise of humanists at their core.

For LLMs to effectively and responsibly leverage this data and become reliable for DH needs, researchers developing these models and the tools that use the models, need to collaborate with data experts. As McGillivray et al. (2020) point out, there is a need for cross-fertilization of ideas and more communication across the NLP and DH communities. LINC (Brown et al., 2023) is an example of computer science (CS) and humanities practitioners working together to extract knowledge from DH texts in the form of linked data connected to web pages to create machine-readable data that could ultimately enhance LLMs. However, the inability of current systems to handle the ontological nuances of the source data plus the absence of entities from popular knowledge bases (KBs) like

*Contributed equally to this work.

¹<https://doi.org/10.5683/SP3/RCVANO>

Wikipedia and Wikidata (Vrandečić and Krötzsch, 2014) necessitates manual entity linking, ontology mapping, and data validation. There is an opportunity here for NLP developers to better support such projects with systems optimized for cultural heritage data.

To contribute to these efforts, **we create an NLP dataset through a collaborative effort between computer scientists and humanists**. Our dataset, *Orlando (Release)* (Hervieux et al., 2024), consists of 12,627 unique text chunks with over 40,000 entity mentions across four entity types that are manually linked to external entity URIs and annotated with 79 unique inter- and cross-sentence relations. The source texts are biographies of historical writers from a large and representative born-digital humanities corpus created by the Orlando Project (Brown et al., 2022) (Appendix A). These source biographies are originally expressed as XML documents, written and thoughtfully hand-annotated in English by DH scholars, using language demonstrably richer than that found in typical LLM training corpora. We extract our dataset from the source while ensuring a high rate of long-tail entities, and preserving the ontological nuances of the source texts’ entity and relationship annotations, which we augment with manually-confirmed entity URIs (§3). This makes the dataset particularly well-suited for information extraction tasks such as entity linking (EL) and relation extraction (RE), as elaborated in §5.

Orlando’s text complexity (examples in Appendix A.1) makes it an interesting subject of study for what machine-aided tools can process. We conduct a series of linguistic analyses (§4) to show that, compared to other genres of text such as news, encyclopedia, or web pages, the Orlando data is more complex in terms of both lexical and syntactic aspects. In light of that, we test whether the Orlando data is out-of-distribution for state-of-the-art large language models such as Llama 2 (Touvron et al., 2023), using metrics based on Mahalanobis distance (Ren et al., 2023) and kernel density estimation (Kirchenbauer et al., 2024). The test highlights that the Orlando data is evidently further from the distributions of the training corpora of LLMs than baseline corpora. This suggests that LLMs, when used out-of-the-box, may suffer from poorer generalization, lower accuracy and higher aleatoric uncertainty (Baan et al., 2023) when processing complex DH text like that from Orlando.

We hope that this data will encourage NLP tool

developers to embrace the challenges posed by DH texts and seek collaboration with the data experts, leading to research, data, tools and systems that would be valuable across disciplines.

2 Related Work

Our dataset is a unique addition to the important yet disproportionately scarce collection of information extraction datasets derived from humanities texts and created collaboratively by NLP and DH researchers. There are countless datasets created to benchmark information extraction models (Nasar et al., 2022) and many works that perform such benchmark evaluations (Wang et al., 2022; Chang et al., 2024), but they typically lack the humanities perspective. There are exceptions, with examples including but not limited to Menini et al. (2022)’s information extraction benchmark relevant to cultural historians interested in textual descriptions of smells in historical documents; Delmonte and Busetto (2023)’s investigation of BERT’s limitations when applied to linguistically complex Italian poetry; Pedinotti et al. (2021)’s diagnostic dataset and evaluation of transformer-based language models on generalized event knowledge; and Bamman et al. (2020)’s challenging coreference resolution dataset for literary texts. These works focus on other genres of text than that of Orlando, and our domain allows us to provide hand-curated cross-database annotations for entity mentions, which is crucial for the evaluation of EL. Compared to works that evaluate BERT and other specialized models, we focus on the suitability of LLMs for humanities-related information extraction tasks. This is critical as LLMs are becoming the status quo for many NLP tasks (Chang et al., 2024) and LLMs are often used with a different paradigm: zero-shot prompting rather than fitting to the target domain.

There are many other valuable datasets coming out of the humanities^{2,3,4}. Major differences between these datasets and ours are that most of these projects release their entire research corpora as raw text with humanities research as the target task rather than information extraction or LLM benchmarking or fine-tuning. Our approach was to look at a prominent humanities dataset with challenging language, consult with DH scholars to understand

²<https://rutgersdh.github.io/dh-sources/>

³<https://humanitiesdata.com/resources>

⁴<https://melaniewalsh.github.io/>

what level of information was important to keep, and then selectively sample it with the intent of keeping difficult chunks that contain many diverse entities and relations.

Our text analysis of Orlando draws on work in evaluating text readability (Crossley et al., 2011; Lu, 2010), but our work deviates as we apply readability measures to compare corpora used in NLP models. We use the popular Flesh-Kincaid grade level (Kincaid et al., 1975) which suits our chunk-level data compared to other metrics like Coh-Metrix (Graesser et al., 2004) which requires paragraph statistics and discourse coherence. See Lu’s (2014) work for a corpus linguist’s review of computational corpus analysis. Our work interrogates whether a corpus is out-of-distribution of an LLM’s training corpus, picking the best-performing indicators in recent discussions (Ren et al., 2023; Yauney et al., 2023; Kirchenbauer et al., 2024).

3 Creating the Orlando Dataset

Derived from the original Orlando XML documents, we release a simplified and easily machine processable JSON dataset, Orlando (Release). Through this collaboration with the data experts, we simplify the complexly nested embedded annotations into an easy-to-use benchmark, without abstracting the nuance of the original entity and relation types.

According to the license, we can release 10% of the Orlando textbase. Instead of uniformly sampling from all sentences or entire documents, we release text chunks of up to 4 sentences each that capture valuable cross-sentence relationships and helpful context for coreference resolution and EL tasks. We filter out text chunks containing too few entities or relations to ensure a high density of useful text. When our extractions come from overlapping chunks, we merge smaller ones into larger ones. We select the included text chunks randomly but with constraints to keep the original frequency distribution for relations and to prioritize the inclusion of person mentions with external entity links. Our sampling process does not alter the distribution of data as it is uniform sampling in a stratified fashion that preserves the long tail distributions of relation and entity types.

Orlando (Release) has 12,627 unique text chunks with over 40 thousand entity mentions across four entity types, with the majority being person mentions. Table 1 lists entity mention counts by type

	Mentions	Entities
person		
bio subjects	14,168	1,389
bio subjects with links	14,122	1,379
others	10,627	6,257
others with links	6,951	3,145
organization	2,910	1,466
place	11,638	4,785
creative work	1,127	928

Table 1: Mention and unique entity count for each entity type in Orlando (Release). Place and creative work types were not de-duplicated so entity count is the number of unique mentions. “with links” rows are subsets of the row directly above.

and presence of external entity links, and breaks down person mentions into the primary subjects of the biographies and other mentioned persons.

Compared to typical RE benchmarks with few broad relations, our dataset contains 79 unique relations, 30 of which are present in at least two contextual categories. The full list of relations and categories with frequency statistics are in the Appendix B Tables 8, 9, and 10.

3.1 Source Textbase

The original Orlando documents are densely annotated XML biocritical profiles of authors (*biographies*). Tags are applied on the word level to identify and add context to entities and concepts, and on the sentence or paragraph level for contextual themes and relations. Figure 1 presents an example. The annotations signal what is most relevant to the domain researchers, which means that not all possible entities and relations are tagged. The data is unique in that the included annotations are extremely detailed, as we discussed in noting the wide range of relations.

Each biography focuses on one person, who we refer to as the biography subject. The biographies are sectioned in two: (1) their birth, death, and the people, places, and activities in between; (2) their writing and its reception. We sample only from the first to prioritize capturing relations between persons without the added complexity of written and often fictional works.

Entity Tags There are four entity types explicitly tagged in the XML documents that we include in our dataset: *person*, *place*, *organization*, and *creative work*. Pronouns are not flagged by annotators

```

<HEADING>Marriage</HEADING>
<FAMILY>
  <MEMBER RELATION="HUSBAND">
    <MARRIAGE>
      <CHRONSTRUCT RELEVANCE="COMPREHENSIVE">
        <DATE VALUE="1834-09-24">24 September 1834</DATE>
        <CHRONPROSE>
          <NAME STANDARD="Adams, Sarah Flower"
            REF="...d681ef">
            Sarah Flower
          </NAME> married
          <NAME STANDARD="Adams, William Bridges"
            REF="...9f83e2">
            William Bridges Adams
          </NAME>,
          <JOB>engineer</JOB> and
          <JOB>inventor</JOB>, at
          <PLACE>
            <PLACENAME REG="St John's Church, Hackney">
              St John's parish church</PLACENAME>,
            <SETTLEMENT CURRENT="London">Hackney</SETTLEMENT>
            <REGION REG="Middlesex"/>
            <GEOG REG="England"/>
          </PLACE>,
        </CHRONPROSE>
      </CHRONSTRUCT>
    </MARRIAGE>
  </MEMBER>
</FAMILY>

```

(a) The original Orlando documents are XML in which text is embedded within deeply-nested relation and entity tags.

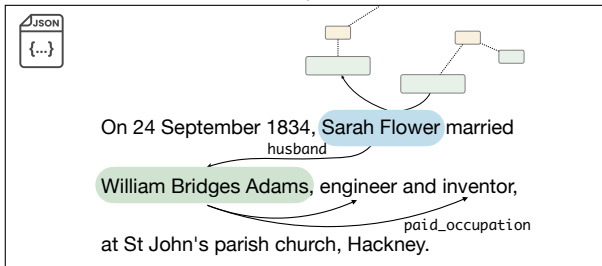
Sarah Flower Adams

22 February 1805 - 14 August 1848

Marriage

24 September 1834
Sarah Flower married William Bridges Adams, engineer and inventor, at St John's parish church, Hackney.³³

(b) Author profile corresponding to the XML in Figure 1a.



(c) We release JSON documents with the same information as the XML but with text extracted and cleaned and lists of relations and entities enhanced with external identifiers.

Figure 1: Representation of an Orlando text chunk displayed as its source XML document (a), its published form on the web (b), and its extracted form in our released dataset (c).

and thus do not appear as mentions in our data. However, there are cases where a person’s relationship to another is used as the mention. For example, “Elizabeth Singer Rowe returned to Frome to live with her father” contains the mention “father”. We include such mentions because a human annotator could confirm a match using the context, so a sophisticated EL method may also be able to.

Relation Tags The XML relation tags indicate how something in the text relates back to the biography subject, making them the subject of all extracted relations. There is, therefore, no specific text span to connect them to a given relation. The only exception is that the occupations of family members are explicitly tagged.

Rather than tagging specific verbs to represent explicit relations, the annotators tag multi-sentence and sometimes multi-paragraph chunks with specific categorical terms⁵. For each category, there are certain nested tags that we use to extract relations. For example, within a <FAMILY> tag, there can be a nested <MARRIAGE> element, within which the first <NAME> element represents the biography subject’s spouse, and within <DEATH>, we may find <DATE>, <CAUSE>, and <PLACE> with details of the biography subject’s death.

⁵<https://orlando.cambridge.org/index.php/about/tag-diagrams>

3.2 Extracted Dataset

Pre-processing We apply an automated text-cleaning step before extracting our dataset from the XML, correcting typographic whitespace errors and integrating dates at the starts of sentences rather than as headings. Originally, the biography subjects are mentioned with project-specific acronyms, which we replace with full names as defined by the annotators. As such, a subject is always mentioned with the same text, except for female subjects when called their birth name early on and their married name later.

Finding Entity Links 20 annotators with backgrounds in humanities and CS⁶ manually searched, using OpenRefine (Delpeuch et al., 2024), for external identifiers for a subset of the over 27,000 unique person entities. To get a broad sample, the review began with the first 8,500 persons by alphabetical order, then 3,240 remaining persons with the highest mention count across all biographies. All biography subjects had been previously searched for by earlier annotators who also found matches for many places in GeoNames (Unxos, 2013).

We instructed annotators to choose one match per entity from VIAF (Tillett, 2002), Wikidata, or Getty ULAN (Harpring, 2010). The Orlando Project leads deemed those sources useful for creat-

⁶The annotators were undergraduate students enrolled in humanities and CS programs and were paid employees of the authors’ universities.

ing meaningful linked open data. Using any available context from the biographies or the web, annotators confirmed matches when they were “definite” or “reasonably certain” on our four-point scale. These scores required multiple pieces of evidence, such as matching birth and death dates, titles of written works, or family members. When annotators could not confirm a match, either because of inadequate evidence or absence of a viable candidate, our data specifies “unable to confirm match.” If an entity was not reviewed by our annotators, we mark it as “match not searched for.” Using the confirmed matches, we query Wikidata’s SPARQL endpoint to get equivalent URIs across the three sources and Wikipedia. All found *links* are included in our dataset to facilitate benchmarking systems that use different KBs.

4 Corpus Comparison

We compare Orlando with corpora of varying genres to determine its complexity for human readers and automated processing.

4.1 Baseline Corpora

We select baselines by two criteria. They spread across diverse genres, including news, encyclopedias, and webpages. They also represent the typical corpora used in training LLMs to provide a more accurate projection of the difficulty of Orlando for LLMs. Each corpus is pre-processed using the same pipeline as Orlando (detailed in Appendix C.1). The corpora are:

C4 Common Crawl⁷ is a large corpus of webpages, reflecting the proportions of different textual content available on the Internet. We use the derived C4 dataset (Raffel et al., 2020a), a cleaned version of Common Crawl that only contains English webpages, as it is the backbone training corpus for many LLMs (Raffel et al., 2020b; Chalkidis et al., 2022; Groeneveld et al., 2024).

CC-News We use the subset of CC-News (Nagel, 2016) prepared by Liu et al. (2019) using *news-please* (Hamborg et al., 2017), which is a dataset of 708,241 English-only news articles extracted from Common Crawl. It is part of the mixture of training corpora of smaller scale language models such as RoBERTa (Liu et al., 2019).

Corpus	FKGL	Avg. Entities
C4	9.56	1.13
CC-News	9.66	1.88
Wikipedia	11.75	2.84
Simple Wiki	8.93	2.16
Orlando (Full)	11.47	2.40
Orlando (Release)	11.90	3.15

Table 2: Flesch-Kincaid Grade Level (FKGL) and average number of entities per sentence of the corpora.

Wikipedia The English Wikipedia is a large encyclopedia that is also widely used as a training corpus for a full spectrum of language models as summarized by Alshahrani et al. (2023).

Simple Wiki As a reference point for text complexity, we include Simple English Wikipedia. It is a version of Wikipedia that is written in simple English and is intended for people with different language proficiency levels.

Orlando (Full) As a baseline, our comparison includes the full text of all biographies in the Orlando Project (Brown et al., 2022). The dataset we release, designated as **Orlando (Release)**, is a subset of the full Orlando dataset that only contains high-quality chunks of text satisfying the criteria described in §3.

4.2 Lexical Complexity

Metrics We count the number of entities in each sentence and report the Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) in Table 2. FKGL maps the number of syllables and words in a sentence to the number of years of education required to understand the sentence, and is widely used in the automatic evaluation of text complexity (Alva-Manchego et al., 2019).

Discussion Overall, Orlando is among the most complex corpora in terms of lexical complexity which could pose difficulties for human readers. The similarity between Orlando and Wikipedia is expected, as they share a similar genre with biographical text comprising a large part of Wikipedia. They contain more named entities per sentence than the other corpora. In particular, the distribution of the number of entities in our released subset skews towards the right, with the highest mean.

From an information extraction perspective, the high number of entities per sentence makes Orlando harder to process as it requires more EL,

⁷<https://commoncrawl.org>

coreference resolution, and RE operations. The released subset is selected with a preference to contain sentences with more entities and relations, which makes it more suitable for EL and RE benchmarking and leads to a more challenging dataset.

4.3 Syntactic Complexity

Metrics We use the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010) to analyze the syntactic complexity of the corpora. This widely adopted tool enumerates a list of patterns in a parse tree and produces 13 variables associated with five aspects of syntactic complexity: length, subordination, coordination, overall complexity, and phrasal sophistication. We plot the scores of the corpora concerning each of the five aspects in Figure 2.

Discussion L2SCA shows that Orlando has higher syntactical complexity than the other corpora. It ranks high in all five aspects of syntactic complexity, with the highest scores in length of production units and amounts of coordination. In comparison, CC-News has high number of subordinations, but fewer coordination and shorter production units. Wikipedia has more subordinations than CC-News and the highest ratio of complex nominals, but it has fewer coordination and shorter production units.

4.4 In-distribution Assessment

Metrics There is mounting theoretical (Saunshi et al., 2021) and empirical (Razeghi et al., 2022; Kandpal et al., 2023; Ren et al., 2023; Kirchenbauer et al., 2024) evidence that suggests a positive correlation between the similarity of the distributions of training and test data and LLM’s performance. Therefore, assessing whether a test dataset (Orlando in our case) is in-distribution, i.e. it follows the same distribution of a model’s training distribution, could be indicative of the model’s relative performance on the Orlando dataset.

We adapt two metrics to measure whether Orlando is in-distribution: Mahalanobis distance (MD; Ren et al., 2023) and kernel density estimation (KDE; Kirchenbauer et al. 2024) with respect to the training data distribution. The metrics are shown to be correlated with model’s performance on translation and language understanding respectively. Both methods represent training and test samples in the embedding space. Ren et al. (2023) fits the training data to a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ and computes the squared Mahalanobis

distance $MD(x) = (x - \mu)^T \Sigma^{-1} (x - \mu)$ for each test sample. Kirchenbauer et al. (2024) takes a non-parametric approach and estimates the probability density of each test sample directly from the training data using the approximate KDE algorithm by Karppa et al. (2022).

The two metrics both require access to the training data of a model, while MD also relies on the internal activations of LLMs. The former is generally unavailable except for works from the LLM open-science community such as (Soldaini et al., 2024) and (Groeneveld et al., 2024). The latter is also unavailable for blackbox LLMs such as GPT-4. For MD, we analyze two open-weight LMs: decoder-only Llama-2-7B (Touvron et al., 2023) and encoder-decoder BART-large (Lewis et al., 2020), and assume C4 to be a good approximation of the training data based on the observations in §4.1. For KDE, we use Soldaini et al.’s (2024) open-science replica to approximate frontier LLMs’ training data.

Discussion As shown in Figure 3, both MD and KDE show that Orlando as test data has lower density in LLM’s training data distribution, indicating that Orlando contains more long-tail information (to be discussed in §4.5) and is more likely to be out of the distribution, compared to general webpages, news or Wikipedia articles. While existing research does not establish a clear density threshold for ensuring the acceptable performance of LLMs, the findings indicate a need for extra caution, as the use of LLMs with DH data may lead to relatively degraded performance. As a future direction, we suggest more directly benchmarking LLMs on DH datasets such as Orlando.

4.5 A High Percentage of Long-tail Entities

Thanks to our URI attribution, we find the percentage of Orlando (Full) person entities in common KBs: Wikipedia for notable people, Wikidata as a larger and more diverse KB, VIAF for people with publications which are relevant to Orlando, and Getty ULAN as an example relevant to many humanities texts but less so Orlando.

Table 3 presents the results for 1,434 subjects and 8,510 randomly sampled other people. Unsurprisingly, over 90% of subjects notable enough to have biographies written about them are found in each of the three relevant KBs. However, 50.8%, 41.1%, and 37.5% of the other people could not be found in Wikipedia, Wikidata, and VIAF, respec-

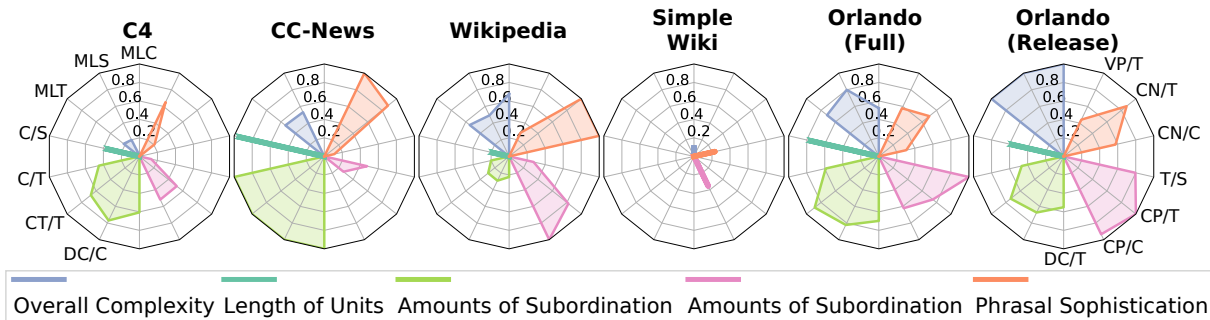
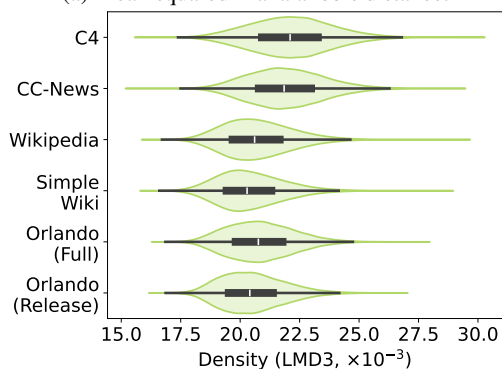


Figure 2: L2SCA values of the six corpora, with each corpus plotted as an individual circle. Within a circle, each polygon represents an aspect of L2SCA and each vertex represents a variable. Starting from 12 o'clock and iterating counterclockwise, the respective aspects of the polygons are listed in the legend. The area of the polygon is proportional to the scores of the aspect. For illustration, the values are normalized by min-max scaling. Definitions of the abbreviated variable names and raw values of the 14 variables of L2SCA (Lu, 2010) in Tables 11, 12 and 13.

	C4	CC-News	Wikipedia	Simple Wiki	Orlando (Full)	Orlando (Release)
Llama-2-7B	0.0	10.2	23.9	30.6	187.2	261.1
BART-large	0.0	9.6	15.7	23.3	102.1	134.6

(a) Mean squared Mahalanobis distance.



(b) Kernel density estimation.

Figure 3: Both MD and KDE show that Orlando is relatively more out-of-distribution. Higher chance of samples being out-of-distribution results in higher MD and lower KDE.

tively by our human annotators – either because the entity was not present or there was insufficient evidence to make a match. These rates highlight that a large percentage of Orlando entities are not considered notable and demonstrate Orlando’s high concentration of long-tail entities (Kandpal et al., 2023).

Kandpal et al. (2023) explore the relationship between question answering performance and the number of documents about an entity in the training data, and report reduced performance in connection with long-tail entities. This raises questions about

how LLMs and associated tools will perform on entity-based tasks with data such as Orlando. There is an opportunity here for LLMs to harness more humanities data to work better for long-tail entities – ultimately reducing historical biases and uplifting historically silenced and overlooked individuals. It also highlights the importance of datasets such as ours so that systems can be evaluated on a mix of popular and long-tail entities.

	Bio Subjects	Others
Wikipedia	93.6%	49.2%
Wikidata	98.7%	58.9%
VIAF	94.3%	62.5%
ULAN	13.2%	8.0%

Table 3: Percentage of unique person entities reviewed by annotators that have matches in each KB. This is a sample of 1434 biography subjects and 8510 others mentioned in Orlando (Full).

5 Exploring Dataset Use Cases

The unique linguistic features of Orlando texts have made them subjects of study in applications like text simplification (Yao et al., 2024). Our dataset enriches the texts with annotations focused on entity and relation mentions, making it well suited for information extraction tasks. We demonstrate the data’s usefulness through off-the-shelf EL and RE systems simple to use without customization.

Entity Linking We use the zero-shot EL system BLINK (Wu et al., 2020, details in Appendix D.1), which uses transfer learning and is potentially useful when applied to the humanities because it should not require training data from the target do-

	Bio Subjects	Others
By Entity	0.89	0.80
By Mention	0.92	0.81

Table 4: BLINK entity linking accuracy on Orlando (Release) using BLINK’s pre-trained Wikipedia model.

main. We link the 13,727 mentions of the 1,307 unique biography subjects and the 5,920 mentions of the unique 2,528 other persons that have confirmed Wikipedia links in Orlando (Release).

Many EL systems consist of an end-to-end pipeline for both recognition (finding mentions to entities) and linking (matching each mention to a database entry). Orlando (Release) enables the evaluation of both steps, but we limit this evaluation to the linking step because BLINK uses a third-party named entity recognition (NER) system. As such, and as is standard in this setting where the system is not able to abstain from making a prediction, we report only accuracy⁸ (Wu et al., 2020; Botha et al., 2020; Hoffart et al., 2011).

Table 4 presents the accuracy broken down by mention and unique entity for each person type. The two rows “By Mention” and “By Entity” refer to two common ways to aggregate results in EL literature. “By Mention” accuracy is micro-averaged as in Hoffart et al. (2011) or the number of correct matches divided by the number of mentions. “By Entity” is the macro-average, calculated as the number of correct matches divided by the number of mentions of entity e , for each entity e in our dataset and then taking the average.

BLINK performs similarly on Orlando subjects to what Wu et al. (2020) report on TACKBP-2010 (0.92 here compared to their best accuracy of 0.945). However, we see a 0.09 to 0.11 point decrease between the notable subjects set and that of the other people, which contains more long-tail entities. It is also important to note that, following Wu et al. (2020)’s problem setup and because of BLINK’s inability to make NIL prediction, we report accuracy only of the entities for which we have confirmed Wikipedia links. This results in artificially inflated scores that are not reflective of the reality of EL on humanities texts. These issues highlight the potential our data has as a challenging entity linking benchmark.

⁸Accuracy and precision are equal in this task setting.

Relation Extraction We use the end-to-end RE system PURE (Zhong and Chen, 2021) on a random sample of 50 text chunks from Orlando (Release), pre-processed as described in Appendix D.2. PURE uses a small set of generic predicates based on those used in the ACE05 dataset (Walker et al., 2006), where a predicate is the connecting term in the subject-predicate-object representation of an extracted relation. On this sample, PURE’s results include six unique predicates, while Orlando (Release) includes 34. Table 5 shows our mapping between Orlando and PURE predicates for the relations that were correctly present in both the Orlando annotations and in PURE’s results for this sample.

PURE	Orlando
General-Affiliation	relocation
Person-Social	brother, husband, interpersonal_relationship, instructor
Physical	habitation, relocation, travel, visit

Table 5: The mapping between PURE predicates and Orlando predicates on the relations that both PURE and Orlando correctly identify on a sample of 50 random Orlando (Release) text chunks. This is only 8 of 34 unique Orlando predicates from this sample that PURE found equivalents to.

Even in this small sample of overlapping predicates, we see PURE abstracting away the valuable specificity in Orlando’s thoughtfully created predicates. Of course there are systems with predicate sets ranging in size and specificity, but we use PURE as a demonstration that our data can be used to evaluate and improve systems across that range.

We manually verify each relation in PURE’s results. Of the 115 relations that PURE extracts, 83% are correct and 65% are both correct and not found in our Orlando annotations. However, the predicates that PURE uses are so high-level that it is challenging to derive meaning from many of the new extractions. PURE only finds 10% of the 174 annotated relations in this sample. This indicates that the detail contained in Orlando poses a significant challenge for such RE systems. Table 6 provides an example, showing the relations that PURE finds for the text in Appendix D.3 that

subject	predicate	object
her	Person-Social	parents
student	Organization-Affiliation	school
her	Organization-Affiliation	school
school	General-Affiliation	Canada
Annie Louisa Walker	social_class	professional but not wealthy rank among the middle classes
Annie Louisa Walker	nationality	English
Annie Louisa Walker	race_colour	white
Annie Louisa Walker	religion	Christian
Annie Louisa Walker	religion	Evangelicals
Annie Louisa Walker	gendered_political_activity	Temperance movement
Annie Louisa Walker	political_involvement	Temperance movement

Table 6: An example of PURE (top) and Orlando (Release) (bottom) relations on the same text sample. PURE abstracts away Orlando’s valuable detailed predicates.

subject	predicate	object
Philip Larkin	school	St John’s College, Oxford
Philip Larkin	subject_studied	English language and literature
Philip Larkin	degree	Honours BA
Philip Larkin	education_companion	Bruce Montgomery
Philip Larkin	education_companion	Kingsley Amis
Philip Larkin	contested_behaviour	Amis and Larkin constituted themselves a two-man parody factory mocking every aspect of university life: the syllabus, the dons, and the aspiring writers such as John Heath-Stubbs.

Table 7: An example of Orlando (Release) relation annotations on text where PURE was not able to identify any relations.

are technically correct, but that lack specificity – even if we were to incorporate coreference resolution into the results. Table 7 shows the detailed and varied relation annotations included in Orlando (Release) for the text in Appendix D.4 on which PURE returns no results.

6 Conclusion

We argue that the impressive results reported by fast-paced NLP research might not reach tools in the DH community due to inherent differences in the kinds of texts they use. In particular, we note that LLMs have been shown to underperform with out-of-distribution inputs compared to experiments where test data comes from the same distribution as the training data (which is the norm in NLP research). While there is currently no machinery to predict the gap in observed performance for a given dataset, we report statistics derived using state-of-

the-art methods that indicate noticeable differences between a corpus derived from a prominent born-digital DH textbase and corpora commonly used as training data in NLP research. We contribute this collaboratively developed dataset and argue for its potential to help close the gap between DH scholars and NLP system developers by serving as a benchmark for existing (and future) tools, as well as a resource for tool development.

Limitations

While we provide extensive statistical analysis using state-of-the-art methods, we consider only one (albeit prominent) DH dataset. Many avenues for future work exist. First, a similar analysis with a larger sample of prominent texts from the DH community, covering a range of genres, is needed. From a tool development point of view, an immediate use of our dataset would be fine-tuning

existing large language models to improve their ability to handle similar texts. We hope that our data will also be used in the evaluation and development of NER, EL, and RE tools that are better equipped to handle rich and complex texts with mentions of rarer entities compared to news and other kinds of texts found on the web. More importantly, we see tremendous potential in using other extant resources from the DH community in the development and evaluation of NLP tools. Such an approach can only lead to more robust tools.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Foundation for Innovation.

We thank LINC⁹, CWRC¹⁰, and the Orlando Project¹¹ for funding students to complete the entity link attribution part of this work, and we are grateful to those students for their contributions.

References

- Saied Alshahrani, Norah Alshahrani, and Jeanna Matthews. 2023. **DEPTH+**: An enhanced depth metric for Wikipedia corpora quality. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 175–189, Toronto, Canada. Association for Computational Linguistics.
- Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. **EASSE**: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.
- Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. **Uncertainty in natural language generation: From theory to applications**. *CoRR*, abs/2307.15703.
- David Bamman, Olivia Lewke, and Anya Mansoor. 2020. **An annotated dataset of coreference in english literature**. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 44–54. European Language Resources Association.
- Chris Biemann, Gregory R Crane, Christiane D Fellbaum, and Alexander Mehler. 2014. **Computational humanities-bridging the gap between computer science and digital humanities (dagstuhl seminar 14301)**. In *Dagstuhl reports*, volume 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc., Sebastopol, CA 95472, USA.
- Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. **Entity Linking in 100 Languages**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.
- Susan Brown, Patricia Clements, and Isobel Grundy. 2022. **Orlando: Women’s writing in the british isles from the beginnings to the present**.
- Susan Brown, Kim Martin, and Asen Ivanov. 2023. **“Linking Out: The Long Now of DH Infrastructures.”**. In Paul Barrett and Sarah Roger, editors, *Future Horizons: Canadian Digital Humanities*, chapter 18. University of Ottawa Press, Ottawa, Ontario K1P 6B9, Canada.
- Susan Brown and John Simpson. 2013. **The curious identity of michael field and its implications for humanities research with the semantic web**. In *2013 IEEE International Conference on Big Data (IEEE BigData 2013), 6-9 October 2013, Santa Clara, CA, USA*, pages 77–85. IEEE Computer Society.
- Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022. **An exploration of hierarchical attention transformers for efficient long document classification**. *CoRR*, abs/2210.05529.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. **A survey on evaluation of large language models**. *ACM Trans. Intell. Syst. Technol.*
- Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024. **Spiral of silences: How is large language model killing information retrieval? - A case study on open domain question answering**. *CoRR*, abs/2404.10496.
- Scott A Crossley, David B Allen, and Danielle S McNamara. 2011. **Text readability and intuitive simplification: A comparison of readability formulas**. *Reading in a Foreign Language*, 23(1):84–101.
- Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. **Bias and unfairness in information retrieval systems: New challenges in the LLM era**. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6437–6447. ACM.

⁹<https://lincproject.ca/>

¹⁰<https://cwrc.ca/>

¹¹<https://orlando.cambridge.org>

- Rodolfo Delmonte and Nicolò Busetto. 2023. [Stress test for BERT and deep models: Predicting words from italian poetry](#). *CoRR*, abs/2302.09303.
- Antonin Delpeuch, Tom Morris, David Huynh, Weblate (bot), Stefano Mazzocchi, Jacky, Thad Guidry, elebitzero, Owen Stephens, Isao Matsunami, Iain Sproat, Albin Larsson, Silvério Santos, allanaaa, kushthede, Sandra Fauconnier, Ekta Mishra, Martin Magdinier, Antoine Beaubien, Lu Liu, Joanne Ong, Fabio Tacchelli, Florian Giroud, Allan Nordhøy, Luca Martinelli [Sannita], Elroy Kanye, Mathieu Saby, and Lisa Chandra. 2024. [Openrefine/openrefine: 3.8.2](#).
- Catherine D’Ignazio. 2021. [Outlier](#). In Nanna Bonde Thylstrup, Daniela Agostinho, Annie Ring, Catherine D’Ignazio, and Kristin Veel, editors, *Uncertain archives: Critical keywords for big data*, chapter 40. MIT Press.
- Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-matrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.
- Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muenighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. [Olmo: Accelerating the science of language models](#). *CoRR*, abs/2402.00838.
- Felix Hamborg, Norman Meuschke, Corinna Breitingner, and Bela Gipp. 2017. [news-please - A generic news crawler and extractor](#). In Maria Gäde, Violeta Trkulja, and Vivien Petras, editors, *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science, ISI 2017*, volume 70 of *Schriften zur Informationswissenschaft*, pages 218–223. Verlag Werner Hülsbusch, Berlin, Germany.
- Patricia Harpring. 2010. Development of the getty vocabularies: Aat, tgn, ulan, and cona. *Art Documentation: Journal of the Art Libraries Society of North America*, 29(1):67–72.
- Natalie Hervieux, Peiran Yao, Susan Brown, and Denilson Barbosa. 2024. [Language Resources From Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs](#).
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. [Robust disambiguation of named entities in text](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Matthew Honnibal, Ines Montani, Sofie Van Lan-deghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. [Large language models struggle to learn long-tail knowledge](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.
- Matti Karppa, Martin Aumüller, and Rasmus Pagh. 2022. [DEANN: speeding up kernel-density estimation using approximate nearest neighbor search](#). In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event, volume 151 of Proceedings of Machine Learning Research*, pages 3108–3137. PMLR.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
- John Kirchenbauer, Garrett Honke, Gowthami Somepalli, Jonas Geiping, Katherine Lee, Daphne Ippolito, Tom Goldstein, and David Andre. 2024. [LMD3: Language model data density dependence](#). In *First Conference on Language Modeling*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Xiaofei Lu. 2010. [Automatic analysis of syntactic complexity in second language writing](#). *International journal of corpus linguistics*, 15(4):474–496.
- Xiaofei Lu. 2014. *Computational Methods for Corpus Annotation and Analysis*. Springer Dordrecht, DORDRECHT, Netherlands.

- Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. [Reference-less quality estimation of text simplification systems](#). In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.
- Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz Fabo. 2020. [Digital humanities and natural language processing: “je t’aime... moi non plus”](#). *Digital Humanities Quarterly*, 14(2).
- Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke van Erp, Inger Leemans, Pasquale Lisena, Raphaël Troncy, William Tullett, Ali Hürriyetoglu, Ger Dijkstra, Femke Gordijn, Elias Jürgens, Josephine Koopman, Aron Ouwerkerk, Sanne Steen, Inna Novalija, Janez Brank, Dunja Mladenic, and Anja Zidar. 2022. [A multilingual benchmark to capture olfactory situations over time](#). In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, LChange@ACL 2022, Dublin, Ireland, May 26-27, 2022*, pages 1–10, Online. Association for Computational Linguistics.
- Sebastian Nagel. 2016. [CC-News](#).
- Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2022. [Named entity recognition and relation extraction: State-of-the-art](#). *ACM Comput. Surv.*, 54(1):20:1–20:39.
- Terhi Nurmikko-Fuller. 2023. *Linked Data for Digital Humanities*. Routledge.
- Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. [Good applications for crummy entity linkers?: The case of corpus selection in digital humanities](#). In *Proceedings of the 13th International Conference on Semantic Systems, SEMANTiCS 2017, Amsterdam, The Netherlands, September 11-14, 2017*, pages 81–88. ACM.
- Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. [Did the cat drink the coffee? challenging transformers with generalized event knowledge](#). *CoRR*, abs/2107.10922.
- Andrew Piper, Richard Jean So, and David Bamman. 2021. [Narrative theory for computational narrative understanding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 298–311. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. [C4: Colossal clean crawled corpus](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. [Impact of pretraining term frequencies on few-shot numerical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. [Out-of-distribution detection and selective generation for conditional language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, Kigali, Rwanda. OpenReview.net.
- Nipun Sadvilkar and Mark Neumann. 2020. [PySBD: Pragmatic sentence boundary disambiguation](#). In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.
- Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2021. [A mathematical exploration of why language models help solve downstream tasks](#). In *International Conference on Learning Representations*.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [Semeval-2020 task 1: Unsupervised lexical semantic change detection](#). *arXiv preprint arXiv:2007.11464*.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxin Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hananeh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. [Dolma: an open corpus of three trillion tokens for language model pretraining research](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.
- Efstathios Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *J. Assoc. Inf. Sci. Technol.*, 60(3):538–556.
- Barbara Tillett. 2002. [A virtual international authority file \(viaf\)](#). In *Record of a workshop on Authority Control among Chinese Korean and Japanese Languages (CJK Authority 3)*, pages 117–139.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

- Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction.
- Unxos. 2013. [GeoNames](#).
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.
- Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. [Measure and improve robustness in NLP models: A survey](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4569–4586, Online. Association for Computational Linguistics.
- Donald J Waters. 2023. The emerging digital infrastructure for research in the humanities. *International Journal on Digital Libraries*, 24(2):87–102.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. [Scalable zero-shot entity linking with dense entity retrieval](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6397–6407, Online. Association for Computational Linguistics.
- Peiran Yao, Kostyantyn Guzhva, and Denilson Barbosa. 2024. [Semantic graphs for syntactic simplification: A revisit from the age of LLM](#). In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 105–115, Bangkok, Thailand. Association for Computational Linguistics.
- Gregory Yauney, Emily Reif, and David Mimno. 2023. [Data similarity is not enough to explain language model performance](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11295–11304, Singapore. Association for Computational Linguistics.
- Zexuan Zhong and Danqi Chen. 2021. [A frustratingly easy approach for entity and relation extraction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 50–61, Online. Association for Computational Linguistics.

A The Orlando Project

The Orlando Project is an ongoing experiment in digital literary history that began in 1995. Its flagship output is a regularly updated online “textbase”, *Orlando: Women’s Writing in the British Isles from the Beginnings to the Present*. As of 2023, the textbase comprises 1444 biocritical profiles of authors from 612 BCE onward, 1261 of them women, contextualized by more than 29,361 free-standing events. 2,995,455 semantic tags annotate its 9,043,111 words with structured references to 37,374 unique persons, 8,696 organizations, 12,114 place names, 47,067 titles, and 30,441 bibliographic sources, as well as embedding relationships among them.

The textbase data has been used for analysis, visualization, and interface design research; its content has fed other DH projects in women’s writing; and its XML schema has served as a foundation for similar projects in the Canadian Writing Research Collaboratory¹².

Few born-digital DH resources feature such extensive annotation, since hand-annotation is costly. However, *Orlando* is representative of much DH work in being organized around profiles of significant individuals that refer to other related entities, and in using complex, nuanced language. Linking entities is a key component of DH infrastructure (Waters, 2023). More efficient and accurate EL for text such as *Orlando*’s would provide immense benefits to DH scholars wishing to enhance their data for publication or analysis, and relationship extraction would provide even further value. Suitably packaged fine-tuned LLMs better equipped to deal with the long and elaborate sentences found in *Orlando* would be equally welcome by the DH community.

¹²<https://cwrc.ca/>

A.1 Illustrating Orlando’s Complexity

The density of facts and complex sentence structures in Orlando make it a valuable DH research tool and present an interesting and potentially quite challenging dataset for NLP systems trained on simpler text.

Sentences contain lists of people with multiple parenthetical clauses and nested relations:

“Dora Carrington formed a lively group (the Wild Group, as they were known at the Slade) with women she remained in close contact with for many years, including Dorothy Brett (later the Honourable), Barbara Hiles (later Bagenal), Ruth Humphries, and Alix Sargent-Florence (the daughter of painter Mary Sargent-Florence and later the wife of James Strachey).”

It is often ambiguous, even to a human reader, as to which relationship is referring to which entity:

“One of her sisters and a niece, Horatia Katherine Frances Gatty (later Eden) and Christabel Maxwell, published writings about her.”

There are multi-step person relations with an unnamed mother in the middle:

“Rosina’s mother’s uncle, Sir John Doyle, was Lieutenant Governor of Guernsey at this time.”

With a high count of meaningful clauses per sentence:

“Louisa Baldwin’s mother, a Welsh-woman born Hannah Jones, was George Macdonald’s second wife.”

“Her mother, born Ann Bee, died on 5 October 1766, and a widowed aunt, another Cassandra, came to keep house for the family.”

B Understanding the Benchmark Dataset

Here we describe and contextualize the fields present in the Orlando (Release) JSON dataset.

Entities For each text chunk, we list entity mentions under *entities* with their *start* and *end* offsets using utf-8 encoding. For each *mention*, we include all text tagged by the annotators, as well as contextual information they added as attributes. This includes *full_name*, which for persons and organizations is a more explicit name or a reformatted name, while for places it is typically the name of the encompassing region. Person and organization mentions have manually deduplicated internal Orlando identifiers, *id*.

For each person mention, we indicate if the associated entity is the primary subject of an Orlando biography through *biography_subject*. This does not necessarily indicate the source document of a text chunk as the subject of one biography could be mentioned in another biography. It can be used as one indicator of a person’s notoriety and allows for separate analyses of the writers and the people connected to their lives.

Relations We use the subject-predicate-object formation to represent extracted relations as triples. Many relations are commutative but we only explicitly list one direction. Table 8, Table 9, and Table 10 detail the relations and contextual categories present in our released dataset.

The *predicate_category* for a relation represents the high level XML tag for the text chunk while *predicate_id* and *predicate_name* represent the specific relation. The *predicate_name* is the relation, while, when available, *predicate_id* is a URI from the CWRC Ontology¹³ that either exactly represents the relation or gives more specific information about the relation. For example, the *number_of_children* relation can have *predicate_id* *cwrc:adoption* to contextualize. Note that the same relation can be present under multiple categories, giving the relation slightly different meaning. For example *subject_studied* can have *institutional_education_context*, *self_taught_education_context*, or *domestic_education_context*.

We do provide utf-8 text spans for the objects of the triples. *object_text* contains the exact mention text of that entity, while *object_id* gives context about that entity from the annotations, when available. For places, the id is either a GeoNames URI for the place or an encompassing region, or a string listing such regions. For people and organizations, it is the de-duplicated Orlando ID. For

¹³<https://sparql.cwrc.ca/>

other types like occupations, id can be an identifier from sources such as the CWRC ontology or Library of Congress Subject Headings¹⁴. For dates, it is standard form YYYY or YYYY-MM-DD, and for nationalities, it is an ISO 3166-2 code.

C Details of Baseline Corpora

C.1 Dataset Version and Pre-processing

The English Wikipedia and Simple English Wikipedia corpora that we use in our comparisons are compiled from recent dumps: `enwiki-20230320` for ordinary English Wikipedia and `simplewiki-20230101` for Simple English.

Every corpus is pre-processed using the same pipeline as Orlando, including sentence splitting using PySBD (Sadvilkar and Neumann, 2020) followed by tokenization and entity recognition using the `en_core_web_sm` model of spaCy (Honnibal et al., 2020). For a consistent comparison across corpora, we count all entities identified by the entity recognition model of spaCy without relying on the manual entity annotations of Orlando.

C.2 Additional Lexical Complexity Statistics

We count the number of characters, tokens, and entities in each sentence and report the distributions in Figure 6 and 5 for each corpus. The lexical complexity score we use is proposed by Martin et al. (2018) which is based on the mean log-rank of word frequencies in a sentence and yields higher scores if more rare words are present in the sentence.

The distribution of sentence length in Orlando, measured by the number of tokens or characters, is similar to that of Wikipedia, with a mean higher than that of C4, CC-News, and Simple Wiki. Sentences in Orlando and Wikipedia contain more syllables and tokens, which is also reflected in the higher FKGL.

C.3 L2SCA Variables

Definitions of the 14 variables of L2SCA (Lu, 2010) are listed in Table 11. For illustration purposes, when plotting the L2SCA variables in Figure 2, we normalize the variables by min-max scaling to the range of $[0, 1]$: suppose x is the vector of raw score of a variable across all corpora, then the normalized score is

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

¹⁴<http://id.loc.gov/authorities/subjects/>

The raw values of the variables are reported in Tables 12 and 13.

D Configuration and Results of Dataset Use Cases

D.1 BLINK Entity Linking Configuration

We use the model that Wu et al. (2020) trained on a 2019 Wikipedia dump. We set the parameter k to 10, according to the authors’ suggestion, so the candidate generation step selects 10 candidates, the ranking step ranks those 10, and we compute accuracy using the one highest ranked prediction. We test three options for the maximum number of contextual tokens: (1) full right and left context within the given text chunk, (2) maximum of 32 tokens on each side of the mention, and (3) maximum of 32 total context tokens as Wu et al. (2020) suggest, but the treatments all had the same results.

D.2 PURE Relation Extraction Configuration

We pre-process the chunks with PySBD (Sadvilkar and Neumann, 2020) for sentence splitting and NLTK (Bird et al., 2009) for word tokenization.

D.3 PURE Relation Extraction Example 1

PURE results are identified with square brackets and Orlando’s with curly brackets.

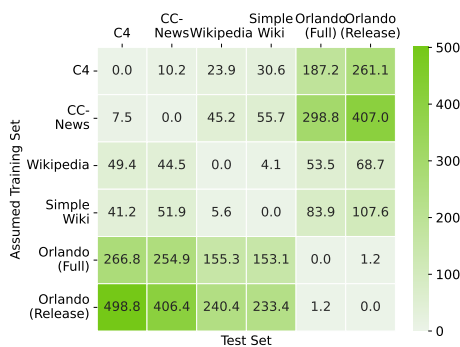
“Coming from a {professional but not wealthy rank among the middle classes}, she seems to have had to contribute to the family income, by teaching and writing, even before [her] [parents]’ deaths. A [student] at [her] [school] in [Canada] described the Walker sisters as very {English}, very dignified, and somewhat exclusive, but... excellent teachers, especially in the departments of history and English literature. Presumably she was {white} and a {Christian} —, one of her verses was appropriated as a hymn by the American {Evangelicals} Dwight L. Moody and Ira Sankey —and she may well have supported the {Temperance movement}.”

D.4 PURE Relation Extraction Example 2

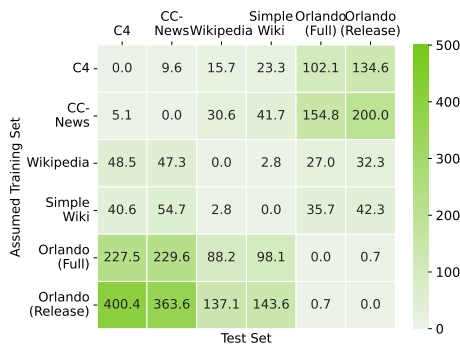
Orlando’s annotations are identified with curly brackets.

“In October 1940 he went up to {St John’s College, Oxford}. He studied

{English language and literature}, and took a {first-class Honours BA} in 1943. Important friendships formed in his undergraduate days were those with {Bruce Montgomery} (who became a highly successful detective-novel writer under the name of Edmund Crispin, and dedicated one of his earliest books to Larkin) and especially the future writer {Kingsley Amis}. {Amis and Larkin constituted themselves a two-man parody factory mocking every aspect of university life: the syllabus, the dons, and the aspiring writers such as John Heath-Stubbs.}”



(a) Llama-2-7B



(b) BART-large

Figure 4: Mean squared Mahalanobis distance of the corpora in comparison to the assumed training data of LLMs.

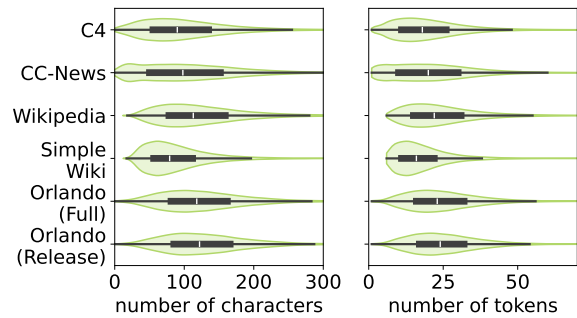


Figure 5: Distributions of the number of characters (left) and tokens (right) in sentences.

predicate_category	count
spatial_context	6818
friends_and_associates_context	5633
cultural_form_context	3275
occupation_context	3101
family_context	2236
birth_context	2210
significant_activity_context	1896
death_context	1816
institutional_education_context	1814
political_context	1676
religion_context	1006
domestic_education_context	826
intimate_relationship_context	728
social_class_context	296
self_taught_education_context	229
nationality_context	141
race_ethnicity_context	113
sexuality_context	107

Table 8: Orlando contextual categories that the relation predicates belong to, with mention counts in our Orlando (Release) dataset.

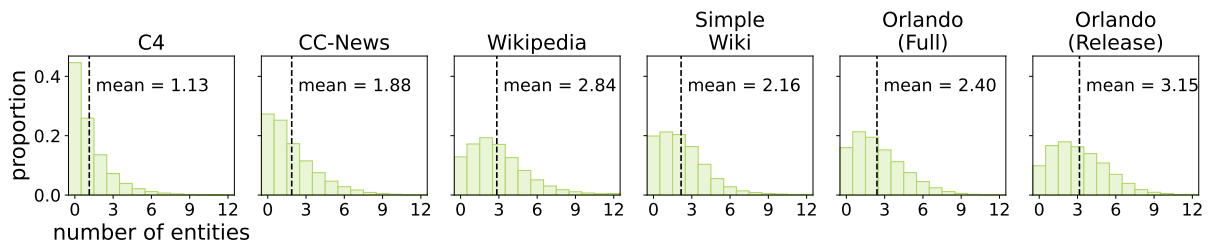


Figure 6: Distributions of the number of entities in sentences of the six corpora.

predicate_name	count	predicate_name	count
interpersonal_relationship	5700	sexuality	137
paid_occupation	2952	brother	136
travel	2336	sister	110
habitation	1846	family_based_occupation	109
relocation	1554	occupation_income	103
subject_studied	1477	son	93
religion	1256	emigration	87
social_class	1011	daughter	85
member_of	974	education_award	55
occupation	940	ethnicity	48
visit	839	cohabitant	47
date_of_birth	803	education_companion	45
place_of_birth	748	other_family	42
date_of_death	695	non_erotic_relationship	42
birth_position	659	intimate_relationship	41
place_of_death	600	contested_behaviour	40
school	594	linguistic_ability	40
nationality	545	grandfather	38
employment	540	uncle	36
activist_involvement_in	525	degree_subject	36
national_heritage	498	wife	35
husband	461	cousin	25
gendered_political_activity	424	grandmother	24
father	399	aunt	21
erotic_relationship	384	native_linguistic_ability	19
cause_of_death	361	spatial_relationship	15
volunteer_occupation	353	forebear	13
political_involvement	336	stepfather	9
mother	307	niece	9
instructor	288	grandson	8
number_of_children	262	nephew	8
political_membership	246	stepmother	7
race_colour	242	granddaughter	5
possibly_erotic_relationship	239	child	3
education_text	179	stepbrother	2
political_affiliation	175	partner	2
burial_location	160	guardian	2
degree	155	stepsister	1
migration	141	stepdaughter	1
geographic_heritage	138		

Table 9: Orlando relation predicates with the mention counts in our Orlando (Release) dataset.

predicate_name	count	predicate_name	count	predicate_name	count
birth_context		friends_and_associates_context		religion_context	
date_of_birth	803	interpersonal_relationship	5608	religion	572
place_of_birth	748	cohabitant	25	member_of	346
birth_position	659	institutional_education_context		social_class	28
cultural_form_context		subject_studied	736	nationality	18
social_class	768	school	585	national_heritage	11
religion	657	instructor	154	activist_involvement_in	7
nationality	434	degree	152	race_colour	4
member_of	432	education_award	52	gendered_political_activity	3
national_heritage	406	education_companion	43	geographic_heritage	3
race_colour	193	degree_subject	36	political_affiliation	3
geographic_heritage	106	education_text	29	sexuality	3
political_Affiliation	44	contested_behaviour	27	ethnicity	2
political_involvement	41	intimate_relationship_context		linguistic_ability	2
sexuality	39	erotic_relationship	384	political_involvement	2
ethnicity	35	possibly_erotic_relationship	239	political_membership	2
activist_involvement_in	29	non_erotic_relationship	42	self_taught_education_context	
linguistic_ability	28	intimate_relationship	41	subject_studied	174
political_membership	27	cohabitant	22	education_text	32
gendered_political_activity	23	nationality_context		instructor	15
native_linguistic_ability	13	nationality	50	education_award	3
death_context		national_heritage	46	contested_behaviour	2
date_of_death	695	social_class	16	degree	2
place_of_death	600	geographic_heritage	8	school	1
cause_of_death	361	religion	8	sexuality_context	
burial_location	160	race_colour	5	sexuality	93
domestic_education_context		member_of	4	activist_involvement_in	3
subject_studied	567	ethnicity	2	gendered_political_activity	3
instructor	119	native_linguistic_ability	2	social_class	3
education_text	118	occupation_context		nationality	1
contested_behaviour	11	paid_occupation	1560	political_Affiliation	1
school	8	employment	540	political_membership	1
education_companion	2	occupation	520	race_colour	1
degree	1	volunteer_occupation	314	religion	1
family_context		occupation_income	103	significant_activity_context	
husband	461	family_based_occupation	64	paid_occupation	1392
father	399	political_context		occupation	420
mother	307	activist_involvement_in	481	family_based_occupation	45
number_of_children	262	gendered_political_activity	391	volunteer_occupation	39
brother	136	political_involvement	289	social_class_context	
sister	110	political_membership	213	social_class	183
son	93	member_of	177	nationality	32
interpersonal_relationship	92	political_Affiliation	123	national_heritage	20
daughter	85	religion	2	member_of	12
other_family	42	race_colour	29	race_colour	10
grandfather	38	national_heritage	15	geographic_heritage	9
uncle	36	social_class	13	religion	9
wife	35	geographic_heritage	12	activist_involvement_in	4
cousin	25	nationality	10	gendered_political_activity	4
grandmother	24	linguistic_ability	9	political_Affiliation	4
aunt	21	ethnicity	8	political_involvement	3
forebear	13	religion	7	political_membership	3
niece	9	native_linguistic_ability	4	ethnicity	1
stepfather	9	member_of	3	linguistic_ability	1
grandson	8	activist_involvement_in	1	sexuality	1
nephew	8	political_involvement	1	spatial_context	
stepmother	7	sexuality	1	travel	2336
granddaughter	5			habitation	1846
child	3			relocation	1554
guardian	2			visit	839
partner	2			migration	141
stepbrother	2			emigration	87
stepdaughter	1			spatial_relationship	15
stepsister	1				

Table 10: Orlando relation predicates with mention counts in our Orlando (Release) dataset. Predicates are repeated in each contextual category (bolded text) in which they appear.

Code	Measure	Definition
MLC	Mean length of clause	# of words / # of clauses
MLS	Mean length of sentence	# of words / # of sentences
MLT	Mean length of T-unit	# of words / # of T-units
C/S	Sentence complexity ratio	# of clauses / # of sentences
C/T	T-unit complexity ratio	# of clauses / # of T-units
CT/T	Complex T-unit ratio	# of complex T-units / # of T-units
DC/C	Dependent clause ratio	# of dependent clauses / # of clauses
DC/T	Dependent clauses per T-unit	# of dependent clauses / # of T-units
CP/C	Coordinate phrases per clause	# of coordinate phrases / # of clauses
CP/T	Coordinate phrases per T-unit	# of coordinate phrases / # of T-units
T/S	Sentence coordination ratio	# of T-units / # of sentences
CN/C	Complex nominals per clause	# of complex nominals / # of clauses
CN/T	Complex nominals per T-unit	# of complex nominals / # of T-units
VP/T	Verb phrases per T-unit	# of verb phrases / # of T-units

Table 11: Descriptions and definitions of variables of L2SCA. The code is used in Figure 2. The table is adapted from Lu, 2010 (pp. 479).

Corpus	MLC	MLS	MLT	C/S	C/T	CT/T	DC/C
C4	11.4955	17.7755	16.8636	1.5463	1.467	0.3563	0.3292
CC-News	11.4438	21.5108	19.6258	1.8797	1.715	0.4279	0.3662
Wikipedia	14.5312	21.079	19.6803	1.4506	1.3543	0.2726	0.2524
Simple Wiki	11.8485	15.5475	14.9764	1.3122	1.264	0.2092	0.2056
Orlando (Full)	13.8285	24.5383	21.1579	1.7745	1.53	0.4033	0.3378
Orlando (Release)	16.0128	26.7895	23.6033	1.673	1.474	0.3696	0.3138

Table 12: Raw scores of the first seven L2SCA variables of the six corpora.

Corpus	DC/T	CP/C	CP/T	T/S	CN/C	CN/T	VP/T
C4	0.4829	0.305	0.4474	1.0541	1.2147	1.7819	1.9977
CC-News	0.6281	0.2355	0.4039	1.096	1.2628	2.1656	2.2858
Wikipedia	0.3419	0.3694	0.5003	1.0711	1.6486	2.2327	1.6942
Simple Wiki	0.2599	0.2829	0.3576	1.0381	1.3151	1.6622	1.4589
Orlando (Full)	0.5169	0.3188	0.4877	1.1598	1.3471	2.0611	1.9369
Orlando (Release)	0.4625	0.3607	0.5316	1.135	1.4648	2.1592	1.8236

Table 13: Continuation of Table 12. Raw scores of the rest of L2SCA variables of the six corpora.

NLP for Digital Humanities: Processing Chronological Text Corpora

Adam Pawłowski

University of Wrocław
pl. Uniwersytecki 1
50-137 Wrocław, Poland
adam.pawlowski@uwr.edu.pl

Tomasz Walkowiak

Wrocław University of Science and Technology
27 Wybrzeże Wyspiańskiego St.
50-370 Wrocław, Poland
tomasz.walkowiak@pwr.edu.pl

Abstract

The paper focuses on the integration of Natural Language Processing (NLP) techniques to analyze extensive chronological text corpora. This research underscores the synergy between humanistic inquiry and computational methods, especially in the processing and analysis of sequential textual data known as lexical series. A reference workflow for chronological corpus analysis is introduced, outlining the methodologies applicable to the ChronoPress corpus, a data set that encompasses 22 years of Polish press from 1945 to 1966. The study showcases the potential of this approach in uncovering cultural and historical patterns through the analysis of lexical series. The findings highlight both the challenges and opportunities present in leveraging lexical series analysis within Digital Humanities, emphasizing the necessity for advanced data filtering and anomaly detection algorithms to effectively manage the vast and intricate datasets characteristic of this field.

1 About Digital Humanities

Digital humanities (DH) today is a broad domain of research and practical applications of various techniques for automatic processing of data, representing linguistics, literary studies, history, art history, cultural anthropology, or archaeology, among others. It can be defined as a system of interrelated resources, functionalities, and cognitive practices created by transferring to the digital realm and creatively expanding the heritage of the humanities that grew out of print culture (Terras et al., 2013; Schreibman et al., 2008; Sinatra and Vitali Rosati, 2014). However, behind the apparent plethora of different descriptions of digital humanities lies the same recurring set of characteristics, among which the most important are: the study of large data sets, the linked open data approach, the extensive use of metadata, the fusion of natural and artificial intelligence, transdisciplinarity, multimedia, an almost

radical empiricism, and last but not least, the dominance of interactive, dynamic infrastructures over the static products of the Gutenberg era, such as articles, chapters, or monographs.

Describing the current state-of-the-art of digital humanities is therefore a task that is all the more difficult because the discipline is undergoing a phase of rapid, dynamic expansion today. Moreover, the efforts of many DH centers focus on digitizing the resources produced by humanity in the past centuries and enriching them with the tools of computational intelligence. The obsolescence of DH definitions, theories, and research practices is hence rapid. Here, we adopt a research perspective in which digital humanities is a strand at the intersection of applied computer science and natural language processing (NLP). This assumption does not contradict the principles of HC, but it exposes the methodological aspects on which NLP's attention is focused, rather than the problem of digitizing and/or sharing resources.

2 Language and NLP vs Digital Humanities

Language has not been a major area of digital humanities in recent years. As the topics of papers submitted to major DH conferences¹ show, interdisciplinary topics dominated there, combining history, art history, cultural anthropology or archival science. The formats processed in DH were, in varying proportions, text, but also image, sound, geographic coordinates (used to formally represent geolocation of objects) and time (dates, hours). However, on a more general level of reflection, natural language preserves its privileged role in the humanities, and this is something that the digital world will not change. After all, language is the primary and universal tool of communication, shaping

¹See <https://adho.org/conference/>, section 'Past Conferences'.

in the human mind a system of representation of the world and modeling the processes of thinking.

Taking the above statement about the privileged role of language in the digital humanities as correct, the potential for collaboration between humanists and language engineers is enormous. NLP applications start with simple text string processing tasks, including segmentation, disambiguation, and morphosyntactic annotation of lexemes, followed by calculating the frequency of any segments. A more advanced level of NLP applications involves generating segment relationships (lexemes or multiword units) that take the form of networks (e.g., wordnets), hierarchical structures (e.g., dendrograms) or point clouds. The next level of NLP application is semantics analysis, which is usually based on distributional relationships of segments (e.g., topic modeling) and/or uses neural networks. A specific class of textual resources is the so-called chronological corpora, which contain samples arranged on a timeline. Below is an example of the application of NLP methods and mathematical statistics to such a corpus.

3 Sequential Data and Chronological Corpora

By textual sequential data, we mean data representing consecutive segments of a corpus on a linear time axis. Representation here means the moment of text production, but is not related to style, genre, age of the writer, or context. Out of the wide spectrum of discourses, the best material to construct chronological corpora are media texts generated by the press, radio, television and all Internet formats. The press is of particular importance here, since it covers long time spans: media institutions in the modern sense have existed in various countries of the world for at least two hundred years. The granularity of the existing resources varies. The earliest texts may have daily or even monthly dates, whereas contemporary news is marked with minute or second accuracy. The chronology of texts can also be discovered in corpora of literary and applied texts, but the granularity in such a situation is at least annual, which makes it necessary to have really large volumes of evenly distributed data.

Sequential text data, which by analogy with time series can be referred to as 'lexical series', obviously have their weaknesses. When extracting electronic text from print, acquiring clean OCR output is a problem. Recordings from radio, television,

and the Internet require transcription supervised by a human. Cleaning and curing the data thus adds significantly to the cost of producing press corpora suitable for chronological analysis. As a result, research on such corpora is not as developed as on large corpora of general language.

An important issue here is to distinguish chronological analysis from diachronic studies. The former deals with changes in the frequency of relevant lexemes at fixed intervals of time, while the latter describes the evolution of language forms over time. Diachronic research can be carried out with quantitative methods, but the nature of the phenomena under study is quite different: it is about change, disappearance and/or appearance of new lexical (rarely morphological or syntactic) forms. An example of a quantitative diachronic research is a study of lexical changes in the Polish language over a period of 600 years (Górski and Eder, 2023). One of the methods for quantitative modeling the dynamics of language change is Pitrovsky's law. This Soviet linguist with Polish roots noted that linguistic changes usually have non-linear patterns, resembling a logistic function (Leopold, 2005; Górski and Eder, 2023).

For the above reasons both approaches should not be confused: they deal with completely different problems, and the reference to time is only an apparent similarity. Bringing them together is all the more difficult because chronological research by NLP methods assumes stable orthography and does not require corpora, derived from long (preferably several hundred years) periods. In contrast, the object of chronological studies are corpora covering shorter periods with stable orthography.

4 Purpose of the Project and Test Data

The purpose of the project is to create a reference processing workflow for textual chronological data and to prepare NLP tools that would serve for lexical series analysis. The first step of the processing flow is to determine the characteristics of the data that are suitable for such analyses and to prepare the test material. The second stage involves the definition of patterns of lexical series, relevant to the needs of the humanities (in particular, linguistics and cultural anthropology), and the preparation of algorithms for extracting such patterns (a priori approach). For some lexemes (represented by lexical series), this stage may include the estimation of trend models and/or stochastic processes if the se-

ries contains periodic oscillations. The third stage involves conducting an unsupervised taxonomy of lexical series, leading to the empirical extraction of semantic classes (a posteriori approach).

As to the first issue, sequences of text samples of the same volume, produced at equal intervals of time (the benchmark example of such a source is the press) should be considered the data best suited for chronological analysis. Additionally, the volume of such samples must be large enough to generate statistically significant lexeme frequencies. Equal sample lengths eliminate in the simplest possible way a troublesome feature of linguistic systems, which is that the dependence of the frequency of the vast majority of lexemes and the volume of the sample is not linear, and, in addition, for each lexeme follows a slightly different curve (in simple terms, if the frequency of a lexeme L in a sample of volume N is L_i , it does not mean that in a sample of $10 * N$ it will be $10 * L_i$ - actually it will be lower due to the constant appearance of new words). Thus, only comparing samples of the same length gives reliable and indisputable results.

As for the sought-after patterns in lexical series, they duplicate and minimally extend the patterns identified in classical time-series analysis, that is, stable trend and periodic oscillations (Box and Jenkins, 1976). In addition, random series and anomalies (“catastrophes”) as non-deterministic patterns are also important in lexical series analysis. The first two types of series express either long-term processes of cultural change (trends) or periodic phenomena, driven by natural cycles (weather, agricultural work, seasonal diseases, etc.) and rituals of culture (anniversaries, holidays, cyclical political events). A non-deterministic pattern with significant informational value is an anomaly, i.e., a sudden jump in the value of a series, caused by some sudden event (the death of a well-known figure, a natural disaster, a change in the name of a great city, etc.). Interestingly enough, in the context of language and culture, researchers may also be interested in random lexical series.

The third stage of the lexical series processing flow potentially includes two modules. If periodicity of the series is detected, a process model (AR, MA, ARMA, ARIMA) can be estimated. However, in the digital humanities, the usefulness of such models is low and it is also difficult to interpret their parameters. As for trend estimation, only working on large datasets covering long periods

gives reliable results. On the other hand, a much more interesting and so far unused approach in this respect is the taxonomy of series. Each lexical series can be treated as a vector representing a point in some multidimensional space. Thus, all lexical series can be projected into this space and can, after dimension reduction, generate human-readable dendrograms or point clouds. It can be expected that such a taxonomy will not be very transparent since placing tens or hundreds of thousands of points in a single space produces a result that is opaque and difficult to interpret. Nevertheless, lexical series (and thus lexemes) with similar frequency characteristics are likely to form at least some visible clusters.

The above processing workflow was tested on the ChronoPress corpus, which represents 22 years of Polish press from the period 1945-1966². This corpus has a volume of ca. 24 millions of segments, evenly distributed by year and month. The texts are lemmatized and chronologically annotated. One year is represented on average by 1,098,526 segments with a standard deviation of 69,843, and the average volume of monthly segments is 91,544 with a standard deviation of 6,201. The volume of the corpus thus allows for annual and monthly granulation. Here, monthly granulation was used, which made it possible to generate lexical series (or vectors) of 264 (=22*12) units in length.

5 Research Methods

In the case of linguistic data, the study of time series consists of two modules. The first comprises NLP tools necessary to generate the input data from the corpus, and the second includes numerical methods of time-series analysis and taxonomy. As for the NLP methods module, the corpus was lemmatized using the WCRFT2 tagger for Polish (Radziszewski, 2013). Each sample includes publication data, allowing us to calculate the number of lexeme occurrences in each month by summing the lexeme numbers obtained from samples published in that month, obtaining time series.

In terms of time series analysis, we rely on the standard model of Box and Jenkins (Box and Jenkins, 1976), exposing, however, the specifics of text corpora. The standard model assumes that any time series consists of a trend, periodic oscillations, and noise. The processing flow includes: (1) identifying the trend and cutting it off from the data,

²<https://chronopress.clarin-pl.eu/>

thus making the series stationary; (2) identifying periodic oscillations by calculating the autocorrelation function (ACF) and partial autocorrelation function (PACF). The shape of the ACF and PACF functions allows one to choose the optimal type and order of the model - autoregressive, moving average, or mixed (ARMA, ARIMA). After cutting off the trend and periodic oscillations from the time series, the residual series that remains should meet the criteria of white noise, and the percentage of explained variance indicates the contribution of the deterministic component to the series.

However, from the perspective of the digital humanities and text processing, some of the functionalities of the standard Box and Jenkins model are not as useful as in economics or engineering. This model, as the title of the cited work indicates (Time series analysis: forecasting and control), was created to predict and/or steer processes. So, while one can understand an economist trying to predict in advance the price of some raw material or the exchange rate of a currency, a humanist does not wonder what the frequency of some lexeme in a stream of media texts will be next month. Therefore, the module for modeling stochastic processes in lexical series of the type described here is not particularly important. An autoregressive model identified in a lexical series would at most show the depth of cultural or societal 'memory'. The previous application of this method in linguistics or textual studies confirms this statement, since the object was to explain some linguistic phenomena, not predict them (Pawłowski, 1997; Pawłowski and Eder, 2001; Mikros and Macutek, 2015). For the above reasons, the ACF and PACF functions should be considered key tools for the humanities to identify periodic phenomena in the great mass of data. This situation raises a fundamental challenge for NLP, namely the need to generate and filter data from a corpus. Chronological analysis of a corpus of texts assumes that there are as many time series as there are different lexemes in the corpus, and the task of the researcher is to identify among them those that are for some reason significant. Despite the great cognitive capabilities of human mind, this task is not feasible without the support of NLP. For example, in the ChronoPress test corpus processed here, the initial number of series was close to 100,000, and this is more likely to be the lower limit, since we are talking about a corpus with an average volume. An additional dif-

ficulty is that the anomaly (catastrophe) pattern can involve series that have subsequent values close to zero almost throughout the run but once their value unexpectedly spikes. An example of such a lexeme is 'comet'. Normally, the press does not write about comets, so the word is almost absent from the media discourse. But, like a real comet, it suddenly appears every few or a dozen years and has higher frequencies. Similarly, lexemes with relatively low frequency, and therefore irrelevant in the perspective of big data, can appear rhythmically. Therefore, the low average frequency of a lexeme is not a sufficient criterion for its elimination. The same applies to the variance of the series, which may be too low to indicate interesting cases of anomalies. To overcome these difficulties, we have developed algorithms for automatic series filtering and anomaly detection.

6 Results

6.1 Overview

From the ChronoPress corpus, a total of 99,528 lexical series with monthly granularity were generated, each with a length of 264 units. The initial issue that required attention was the normalization of the data in order to facilitate comparison. We divided each occurrence of a lexeme by the total occurrences of that lexeme. This allowed us to obtain the probability density function of lexeme occurrences over the analyzed time period.

6.2 Linear Regression

In the first set of experiments, we applied linear regression to normalized time series to calculate the slope and the coefficient of determination R^2 for each lexeme. Next, we have planned to identify the lexemes with the highest, closest to zero, and lowest slopes, representing those that were the most rising, flat, and descending, respectively. However, a decision needed to be made regarding the inclusion of all series in the analysis. Empirical evidence suggested that many series deviated significantly from linearity. Therefore, we excluded series with R^2 values less than 0.5, resulting in the retention of 138 lexical series. The results, including the 10 lexemes with the largest slope, the slope closest to zero, and the smallest one, are presented in Table 1. First of all, we can observe that the majority of lexemes (99.86%) are non-linear, with a coefficient of determination smaller than 0.5. Among those assumed to be linear, functional words, such as

rising	flat	descending
West German	my	self-help
specialized	such	rebuilding
television set	just	fascism
currently	a few	rebuild
set	very	fascist

Table 1: Lexemes with the most rapidly increasing, consistently flat, and sharply declining time series among linear ones (i.e., time series with an R^2 greater than 0.5).

old, at the same time, valuable, date, prove, put in, get to know, upbringing, beginning, leave

Table 2: The 10th most stationary lexemes. Those for which their time series yield the smallest p-values for the Augmented Dickey-Fuller test (the p-values for presented lexemes are approximately e^{-30}).

pronouns and adverbs, tend to have a flat shape, with a slope close to 0. The most rapidly declining words are those associated with World War II and the process of rebuilding after the damage of war.

The interpretation of this result is very positive from the point of view of the efficiency of the model. The ChronoPress corpus reflects the events of post-war Poland, where the trauma of World War 2 is very strong right after 1945, but is gradually dying out, replaced in the official propaganda by Cold War events (e.g., the Korean War). In contrast, the behavior of function words is completely different. Their frequency is largely independent of the sample size, which is the reason why they are massively used in stylometry. Our research has shown that they are also immune to the time factor: successive samples of texts on the timeline are stable in this respect. Finally, technology-related lexemes (e.g., television set) are trending upward.

6.3 Autoregressive Model

In order to detect lexemes with seasonal patterns, we applied the Partial Autocorrelation Function (PACF), which helps to determine the order of an autoregressive model. PACF assumes the stationarity of the underlying time series. We used the Augmented Dickey-Fuller (ADF) test to check the stationarity of the time series with a periodic component. Table 2 shows the most stationary lexemes, where the ADF test p-value is the smallest. These lexemes express cultural rituals (such as religious and national celebrations) and cycles of nature.

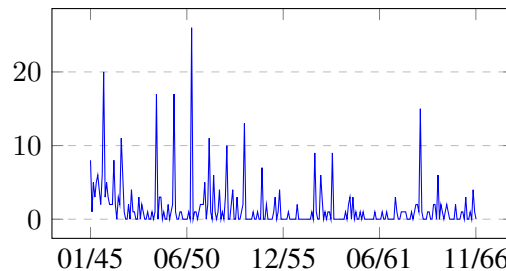


Figure 1: Occurrence of the lexeme 'Lenino' as a time series.

Christmas Eve, New Year, Lenino, August, Barbour, Christmas, April, May, September, Christmas tree

Table 3: The lexemes with the highest values of partial autocorrelation at lag 12 among those assumed to be stationary. The partial autocorrelation values range from 0.66 (for Christmas Eve) to 0.54 (for Christmas tree).

To choose stationary time series, we only consider lexemes with a p-value of ≤ 0.05 . This allows us to reject the null hypothesis, indicating that the time series does not have a unit root and is stationary. Resulting in the retention of 93,039 lexical series. The lexemes with the highest PACF values for lag 12 are shown in Table 3. We can notice some expected events occurring once a year like Christmas Eve, New Year or name of months. However, Lenino looks at first surprisingly (refer to Figure 1 for Lenino time series). But, as the battle of Lenino (12.10.1943) was an important event for communist propaganda, the algorithm found it to be as oscillating as Christmas. The battle of Lenino was the the baptism of fire of Polish troops organized in the USSR, and from 1950 to 1991 it was celebrated as the Polish Army Day³. Of course the event had an important press coverage.

Table 4 presents the lexemes with the highest Partial Autocorrelation Function (PACF) values for a lag of 1. These lexemes are associated with the theme of postwar reconstruction in Poland. The high PACF values, ranging from 0.85 to 0.91, may indicate that these lexeme time series likely follow an AR (1) (AutoRegressive) process. This means that the present value of each series is primarily influenced by its immediate past value.

³https://en.wikipedia.org/wiki/Battle_of_Lenino

reconstruction, occupation, Poland, fascism,
 Polish, democracy, war, UNRRA, destroyed,
 allied

Table 4: The lexemes with the highest values of partial autocorrelation at lag 1 among those assumed to be stationary. The partial autocorrelation values range from 0.91 (for reconstruction) to 0.85 (for allied).

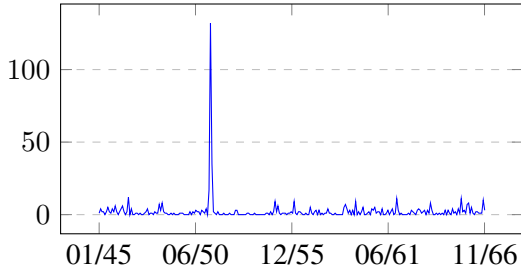


Figure 2: Occurrence of the lexeme 'plebiscite' as a time series.

6.4 Catastrophe Detection

In order to identify sudden spikes in a series, we calculate the difference between the current value of a normalized time series and a moving average with a length of 5. We have defined a catastrophe (anomaly) index as the maximum absolute value of this difference. An issue we needed to address was determining which series should be included in the catastrophe (anomaly) analysis. We decided to exclude series with a low sum (less than 500 occurrences in the entire series). After this exclusion, we were left with 6612 lexical series, which we then analyzed for anomalies. The lexemes with the highest values of the 'catastrophe index' are displayed in Table 5. The values for these lexemes range from 0.17 to 0.07.

An analysis (refer to Figure 2) of the time series for the anomaly of the lexeme 'plebiscite' reveals that it originates from the National Plebiscite for Peace held in Poland in May 1951 under the auspices of the Polish Committee of the Defenders of Peace (Dawid, 2018). The lexeme 'Dzerzhinsky' (peaking in June 1951) is linked to the 25th anniversary of his death and the unveiling of a monument in his honor in Warsaw.

6.5 Clustering and Dimension Reduction

In the next step, we cluster the analyzed lexemes using the normalized time series as vectors. We have utilized the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm (Campello et al., 2015). How-

plebiscite, Dzerzhinsky, senate, ratification
 Communist Party of Poland, coffin, referendum
 rally, constitution, pre-convention, Stalin
 Indochina, Grunwald, Potsdam, capitulation

Table 5: The lexemes with the highest values of the catastrophe index.

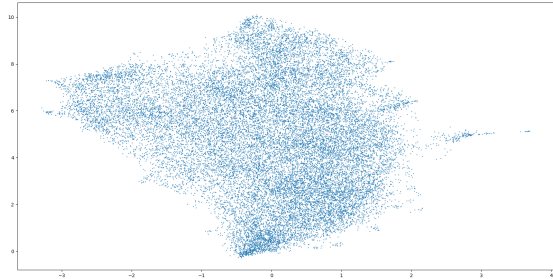


Figure 3: UMAP projection of lexemes - represented by 264 dimensional vectors (default parameters of UMAP method).

ever, it did not detect any distinct groups (only one, large group was identified). Modifying the default values of the HDBSCAN parameters and trying different metrics (the default being Euclidean) did not yield different results. To better understand the issue, we employed the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) technique to reduce dimensions and visualize the similarities in the data. The results are presented in Figure 3. It is clear that UMAP does not reveal any distinct groups in the data, suggesting a relatively equal distribution of vectors in multidimensional space. This is probably why HDBSCAN was unable to identify any well-defined clusters in the data.

Should this result be evaluated entirely negatively? In our opinion, the method described above is not inefficient per se, but it may well be that it performs well when processing data are structured differently (smaller granularity, higher volume of samples, representing years or even decades).

6.6 Similarity of Lexems

A distance metric between vectors defined by normalized time series could be used to identify lexemes that have a similar 'longitudinal shape', i.e. the pattern of their distribution over time line. The results for exemplary lexemes are displayed in Table 6. Some of the results are predictable, such as the similarity between 'harvest' and semantically related terms. However, others, such as the pairing of 'worker' with 'fight' or 'forefront' reveal

war	harvest	flood
Poland	harvesting	simple
government	July	flooded
representative	harvest like	touch
series	July	loss
camp	harvester	water
association	rye	dear
Polish	June	last
area	grain	height
huge	barley	bridge
state	cartload	none

Stalin	Gomułka	worker
leadership	Władysław	female worker
Stalinist	tendency	fight
brilliant	self-gover.	mass
successor	common sense	association
proletarian	of course	fighting
Leninism	environment	people
invincible	opinion	working
Leninist	surely	forefront
leader	demand	factory
generalissimus	view	segment

Table 6: Lexeme similarity results. The lexemes most similar to the exemplar lexemes (those in bold) in order of increasing distance defined over normalized time series vectors.

specific features of the communist propaganda discourse of the post-war period not present in the general language.

7 Conclusions

While traditional time series analysis seems to be a task of fortune-telling (who wouldn't want to know next week's stock prices...), lexical series analysis is akin to looking for a needle in a haystack. In the research presented here, the 'needle' was the time series, containing the trend, periodic oscillations and anomalies, and the haystack was the corpus of 24 million words, divided into 264 sections. In addition, we were looking for lexemes that display similar shapes of time series.

The techniques elaborated during our research allowed us to identify automatically a set of lexemes from the corpus of nearly 100,000 that we found relevant in some way. In particular anomaly / catastrophe detection helps us to pinpoint lexemes that undergo rapid changes in occurrence, with some unexpected cases like "plebiscite" or "Dz-

erzhinsky". Another promising technique involves detecting lexemes with similar time-series patterns. The semantic similarity between words has been a foundational concept in modern NLP, based on deep neural networks and generative models. The Transformers architecture (Vaswani et al., 2017; Devlin et al., 2019) originates from the word2vec method (Mikolov et al., 2013), which creates word representations using large data sets and word occurrences in similar contexts. The method presented here, utilizing normalized time series, shares similarities with word2vec, as it constructs vectors from a large corpus. However, the novelty of our approach is the reliance on co-occurrence in recurring units of time (e.g., years) rather than co-occurrence in the text. The results shown in Table 6 reveal natural recurrence patterns, such as "harvest" and "July," but also provide insights into the communist perspective on the world, identifying semantic clusters typical for the totalitarian propaganda of the communist period. Last but not least, the great advantage of the method developed is that it is language-independent – any chronological (longitudinal) corpora can be processed in this way.

Limitations

The processing workflow presented here was developed and tested on a single corpus of Polish. However, linguistic aspects seem to be the easiest to overcome, as the workflow can be easily extended to other languages by using appropriate, language-specific taggers, such as those available from the spaCy framework (Honnibal et al., 2020). The main limitation of a reliable analysis of chronological corpora is their volume and time distribution. To derive statistical patterns, the corpus must be sufficiently large and balanced in terms of the distribution of analyzed time slots over time. While large corpora of contemporary language are numerous, balanced coverage of long time periods is rare. And the essence of chronological analysis is precisely to describe the "long duration" – far longer than the Internet era – which allows the correct identification of events, processes, cultural phenomena, etc. Practice shows that the most effective method of expanding chronological corpora is to scan the press, which significantly increases the cost of such an activity. Filling in data gaps could be achieved by subsampling underrepresented time periods, but this would result in data loss.

Another limitation is the number of parameters that must be set to use the proposed methods, which can influence the results obtained. These parameters include the minimum value of the coefficient of determination in linear regression analysis and the minimum size of the analyzed time series in the case of anomaly (catastrophe) detection.

Ethics Statement

The data and resources used in this study contain no sensitive data, they are publicly available and have been used in other researches.

Acknowledgements

The work was financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure" (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01.

References

- George Box and Gwilym Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day.
- Ricardo Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data*, 10(1).
- Adriana Dawid. 2018. Organization and running of the national plebiscite for peace in opole voivodeship. *Rocznik Ziemi Zachodnich*, 2018(2).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Rafał L. Górski and Maciej Eder. 2023. Modelling the dynamics of language change: Logistic regression, piotrowski's law, and a handful of examples in polish. *Journal of Quantitative Linguistics*, 30(1):125–151.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.
- Edda Leopold. 2005. Das piotrowski-gesetz. In *Quantitative linguistik – quantitative linguistics. Ein Internationales Handbuch*, page 627–633. de Gruyter.
- Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- George K. Mikros and Ján Macutek, editors. 2015. *Sequences in Language and Text*. De Gruyter Mouton, Berlin, München, Boston.
- Adam Pawłowski and Maciej Eder. 2001. Quantity or stress? sequential analysis of latin prosody. *Journal of Quantitative Linguistics*, 8(1):81–97.
- Adam Pawłowski. 1997. Time-series analysis in linguistics: Application of the arima method to cases of spoken polish. *Journal of Quantitative Linguistics*, 4(1-3):203–221.
- Adam Radziszewski. 2013. A tiered crf tagger for polish. In *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, pages 215–230. Springer Berlin Heidelberg.
- Susan Schreibman, Ray Siemens, and John Unsworth. 2008. *A Companion to Digital Humanities*. Wiley Publishing.
- Michael Sinatra and Marcello Vitali Rosati, editors. 2014. *Pratiques de l'édition numérique*. Parcours numériques. Les Presses de l'Université de Montréal, Montréal. OCLC: 870269451.
- Melissa Terras, Julianne Nyhan, and Edward Vanhoutte. 2013. *Defining Digital Humanities: A Reader*. Ashgate Publishing Company, USA.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Multi-task Framework with Enhanced Hierarchical Attention for Sentiment Analysis on Classical Chinese Poetry: Utilizing Information from Short Lines

Quanqi Du

LT3, Ghent University
Groot-Brittanniëlaan 45, 9000 Gent
Flanders, Belgium
quanqi.du@ugent.be

Veronique Hoste

LT3, Ghent University
Groot-Brittanniëlaan 45, 9000 Gent
Flanders, Belgium
veronique.hoste@ugent.be

Abstract

Classical Chinese poetry has a long history, dating back to the 11th century BC. By investigating the sentiment expressed in the poetry, we can gain more insights in the emotional life and history development in ancient Chinese culture. To help improve the sentiment analysis performance in the field of classical Chinese poetry, we propose to utilize the unique information from the individual short lines that compose the poem, and introduce a multi-task framework with hierarchical attention enhanced with short line sentiment labels. Specifically, the multi-task framework comprises sentiment analysis for both the overall poem and the short lines, while the hierarchical attention consists of word- and sentence-level attention, with the latter enhanced with additional information from short line sentiments. Our experimental results showcase that our approach leveraging more fine-grained information from short lines outperforms the state-of-the-art, achieving an accuracy score of 72.88% and an F1-macro score of 71.05%.

1 Introduction

Classical Chinese poetry generally refers to Chinese poetry composed in certain traditional forms and written in classical Chinese, a special terse, rhythmic and musical version of Chinese that is different from the modern Chinese today and mainly used in classical literature. As one of the earliest forms and most important genres of traditional literature, classical Chinese poetry is a crucial carrier of the thoughts and emotions of ancient Chinese literati.¹ In fact, it is widely believed that “emotion” is the essence of classical Chinese literary culture (Xia, 2021). By investigating the emotion expression in classical Chinese poetry, we are able to have a glimpse of the personal and social issues involved in the poetry, through which we can

¹<https://www.zgbk.com/ecph/words?SiteID=1&ID=389600&Type=bkztb&SubID=683>

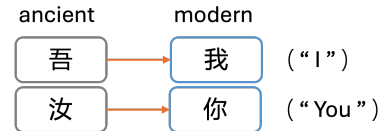


Figure 1: Pronoun changes from ancient Chinese to modern Chinese.

gain more insights into the ancient Chinese culture (Hou and Frank, 2015; Wei et al., 2020; Zhang et al., 2023).

However, compared with the study of emotion in other text genres we encounter in modern life, such as news articles (Nemes and Kiss, 2021) and reviews (Yi and Liu, 2020), there is less research on poetry (Ahmad et al., 2020a), not to mention on classical Chinese poetry (Tang et al., 2020). Actually, it is more challenging to process ancient Chinese texts than modern Chinese text. On the one hand, the classical Chinese has some old vocabulary and phrases that have been replaced or fallen out of usage in modern Chinese. For instance, the pronouns in ancient Chinese are quite different from those in modern Chinese, as shown in Figure 1. On the other hand, the grammar of ancient Chinese is more concise and flexible compared to modern Chinese. As for classical Chinese poetry, it is considered as a more complicated synthesis of imagery and symbolism, allusions, quotations and derivation, and antithesis (Liu, 2022). All these factors make it a difficult task to conduct sentiment analysis in classical Chinese poetry.

To analyse sentiment in classical Chinese poetry, previous studies have explored different methods, for example, constructing sentiment lexicons (Hou and Frank, 2015; Zhang et al., 2023), transferring knowledge from modern Chinese (Zhao et al., 2014), or extracting imagery words (Shen et al., 2019; Su et al., 2023). Although these studies improved the general performance for the task

of sentiment analysis in classical Chinese poetry, by utilizing special words in the poems or drawing upon knowledge beyond the poems, they did not consider the compositional structure of the poems. Usually, a classical Chinese poem comprises several short lines, which may show different emotions, and in return, contribute to the overall emotion expression of the poem. Thus in this paper, for the task of sentiment analysis of classical Chinese poetry, we propose to take the sentiment of short lines into consideration by using a multi-task framework with a hierarchical attention network, which includes the sentiment analysis task of both the overall poem and the short lines of which the poem is comprised. We will show that, by leveraging the sentiment information from the short lines, we can outperform the current state-of-the-art in sentiment analysis of ancient Chinese poetry.

2 Related Work

2.1 Sentiment analysis for computational poetry studies

Sentiment analysis using computational methodologies is receiving increasing attention in literary studies (Kim and Klinger, 2021). In 2012, Kao and Jurafsky applied sentiment lexicons to investigate whether the affect expressed in a poem makes it more beautiful or not. Later, Delmonte et al. (2013) introduced the SPARSAR system for the syntactic, semantic and prosodic analysis of poetry, which also includes a sentiment analysis module. In the comparison of formal and informal texts, Kaur and Saini (2014) found that taking into account genre-specific features helped to improve sentiment analysis performance on formal texts (including poetry). More recently, the advantages of machine learning further promoted the task of sentiment analysis in poetry. Promrit and Waijanya (2017) used Convolutional Neural Networks (CNNs) for category classification and sentiment analysis for Thai poems. Ahmad et al. (2020b) proposed an attention-based C-BiLSTM model to classify poetry texts into different emotional states, while Rajan and Salgaonkar (2020) employed a Naïve Bayes classifier on Konkani Poetry. Similar researches were also carried out on classical Chinese poetry (Tang et al., 2020), Punjabi poetry (Kaur and Saini, 2020) and Latin poetry (Sprugnoli et al., 2022).

2.2 Sentiment analysis for classical Chinese poetry

Although classifying the sentiment in classical Chinese poetry helps to investigate the ancient cultural and emotional life, this task is not as popular as in texts of other fields. Most studies in this field focus on the poems from the Tang Dynasty, which is believed to be the Gold Age of poetry, with over 50,000 poems created and surviving to this day.² On the basis of the Complete Anthology of Tang Poetry, Hou and Frank (2015) proposed a novel graph-based method to create a sentiment lexicon for classical Chinese poetry, with which they analyzed the association of sentiments with different poets and a variety of topics. To take full advantage of deep learning and linguistic knowledge, Zhang et al. (2023) combined supervised sentiment term extraction and classification to incorporate linguistic knowledge into deep learning models for the task of sentiment lexicon construction.

Besides the construction of sentiment lexicons, scholars also investigated extracting more information from the poems. Li and Li (2018) introduced the Frequent Pattern Growth Algorithm with the Term Frequency-Inverse Document Frequency to capture the hidden relationships between each word. (Shen et al., 2019) explored ways to extract sentimental imageries at the levels of both character and word and integrated this information in the task of sentiment analysis. To utilize the imagery words, Su et al. (2023) introduced related visual modality information and obtained better performance in sentiment analysis of classical Chinese poetry.

In addition to the information present in classical Chinese poetry, some scholars also tried to incorporate external knowledge into the task of sentiment analysis. Zhao et al. (2014) introduced the modern translations of ancient texts and transferred this knowledge in the translation to the classical Chinese poems. Liu et al. (2020) incorporated the knowledge of classical Chinese poetry from appreciation/translation annotations into the knowledge graph construction.

Furthermore, more recently, the development of large language models also aided in the promotion of research on classical Chinese literature. In 2021, Tian et al. released AnchiBERT, a pre-trained language model based on the architecture of BERT

²<https://www.zgbk.com/ecph/words?SiteID=1&ID=272482&Type=bkdzb&SubID=808>

and trained on large-scale ancient Chinese corpora. Then, based on BERT-base-Chinese from Google ³ and Chinese-BERT-wwm (Cui et al., 2021), Wang et al. (2022) designed the pre-trained model SikuBERT and SikuRoBERTa specifically for ancient Chinese. There are also other similar models for ancient Chinese, e.g., BERT-ancient-Chinese (Wang and Ren, 2022), GuwenBERT, ⁴ GujiBERT (Wang et al., 2023), and BERT_CCPoem, ⁵ which is particularly designed for classical Chinese poetry.

2.3 Sentiment analysis with multi-task learning and a hierarchical attention network

Traditional sentiment analysis is approached as an independent single task, but recent research has shown that it can also be considered as one of the sub-tasks in multi-task learning. Balikas et al. (2017) incorporated two sentiment analysis tasks (one with a 3-category label set and the other with a more fine-grained 5-category label set) into a multi-task model and demonstrated benefits by jointly learning the two sub-tasks for tweets. Aspect-based sentiment analysis is also a field where it is a common practice to simultaneously model the aspect term extraction and the corresponding sentiment classification (He et al., 2019; Zhao et al., 2023; Wu et al., 2023; Rani and Jain, 2024). Scholars also investigated combining with sentiment analysis other emotion-related tasks, e.g., complaint identification (Singh et al., 2022) and sarcasm detection (Tan et al., 2023). Chauhan et al. (2020) even proposed an all-in-one multi-task framework that incorporates five sub-tasks, including humour, sarcasm, offensive content, motivational content detection and sentiment analysis.

Following the “Attention is all you need” credo from Vaswani et al. (2017), the attention mechanism has been investigated in many variants and the hierarchical attention network (HAN) (Yang et al., 2016) is one of them. HAN contains the attention mechanisms at different levels, such as the word-level and sentence-level, capturing important words/sentences in the task context. It has been shown that HAN performs well with long-sequence information, especially when the information is ordered in a certain structure, for example in document-level classification (Pappas and Popescu-Belis, 2017) and translation (Werlen et al.,

³<https://huggingface.co/google-bert/bert-base-chinese>

⁴<https://github.com/ethan-yt/guwenbert>

⁵<https://github.com/THUNLP-AIPoet/BERT-CCPoem>

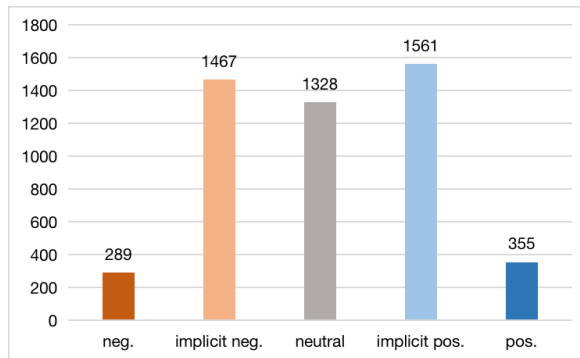


Figure 2: Sentiment label distribution in FSPC, with *negative* and *positive* abbreviated as *neg.* and *pos.* respectively.

2018). HAN is also applied in sentiment-related tasks. Cheng et al. (2017), for example, designed a network with aspect attention and sentiment attention for aspect-level sentiment analysis. Mirroring the structure of the social media content, Cheng et al. (2019) proposed a network with HAN for cyberbullying detection on Instagram. HAN is also used to detect depression from transcribed clinical interviews (Mallol-Ragolta et al., 2019). More recently, Chanaa et al. (2021) exploited HAN for the task of E-learning text sentiment classification.

As mentioned above, previous studies suggest that multi-tasking helps to improve the performance of sentiment analysis, but there is little such research in the field of classical Chinese poetry, which is mainly investigated by constructing auxiliary lexicons or by fusing knowledge from sources beyond the poetry. On the other hand, as classical Chinese poetry is highly structured and information-condensed, we hypothesize it is promising to apply HAN for sentiment analysis of this special genre, which has never been done before. In this paper, we aim to utilize the sentiment knowledge of the short lines to help predict the overall sentiment of the poetry with a HAN-based multi-task framework.

3 Dataset

Building a classical Chinese poetry corpus for sentiment analysis is not trivial, since full comprehension of the poems is one of the most important preconditions for sentiment labelling, which sets high expectations for annotators. Thanks to the work of Chen et al. (2019), we now have a fine-grained sentimental poetry corpus (FSPC), which we are going to utilize in our experiments.

FSPC is composed of 5,000 classical Chinese

lines	sentiment	overall sentiment
诸侯分楚郡 The lords divide the land of Chu	neutral	negative
饮饯五溪春 Feasting and bidding farewell in the spring of Five Rivers	positive	
山水清晖远 The distant beauty of mountains and waters shines clear	neutral	
俱怜一逐臣 All pity this one exiled courtier	negative	

Figure 3: An example of the different sentiment labels at the line level and the overall sentiment of a given poem. Note that the English translation comes from ChatGPT and works only as reference.

poems. Each poem and each line in the poem are manually annotated by experts in Chinese literature with five classes, ranging from *negative* and *implicit negative*, over *neutral*, and to *implicit positive* and *positive*, as shown in Figure 2. Compared with *negative* and *positive*, implicit sentiments refer to the emotions that are suggested or hinted at but not directly stated. Due to the unbalanced label distribution, the implicit sentiments are merged into *negative* or *positive*, respectively, in later experiments. It should also be noted that the sentiment of each line can be different from each other, as shown in Figure 3, with the sentiment of the first line being neutral or implicit while the last line is often aligned with the overall sentiment of the poem (Chen et al., 2019). This suggests each line of the poem may have a varied contribution to the holistic sentiment of the whole poem, and it would be beneficial if we take into consideration the unique sentiment label of each line in the sentiment analysis of the whole poem.

4 Method

To investigate whether the task of sentiment analysis for the whole poem can benefit from the introduction of short line information, we use a framework that fine-tunes a pre-trained model with labels of both the lines and the overall poem, as shown in Figure 4.

Data pre-processing Each poem consists of 4 lines, and along with the poem, sentiment labels of both the lines and the overall poem are used for the pre-trained model fine-tuning. The use of “|” as separators between short lines in FSPC makes it easy to identify the line border.

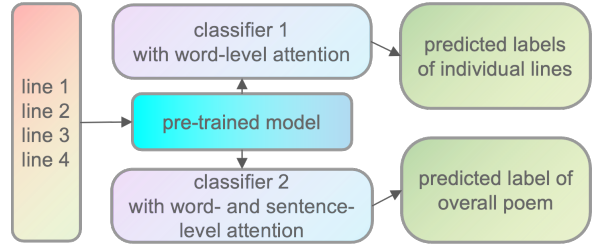


Figure 4: Framework to fine-tune the pre-trained model to predict labels of individual lines and the overall poem.

Feature extraction and fine-tuning The pre-trained model is used both in the feature extraction and fine-tuning stage. In the former case, the architecture of the pre-trained model is preserved, taking in the texts and outputting the feature vectors for the next stage. In the fine-tuning stage, two linear transformation layers are added at the end of the pre-trained model as classifiers for sentiment in the short lines and the overall poem, respectively. Both classifier 1 and classifier 2 use word-level attention, and sentence-level attention is also applied in classifier 2. For both the lines and the overall poem, the cross-entropy loss between the predicted logits and the true labels is calculated. The two loss values are then added to get a combined loss, based on which backpropagation is performed to compute gradients for model parameter updates.

It should be noted that for the short lines, each of them is independently encoded, and the encoding of the whole poem and the short lines are conducted separately.

Enhanced sentence-level attention To exploit the short line label information in the sentiment prediction task of the overall poem, additional information derived from the line labels is integrated into the sentence-level attention scores.

Output evaluation For the evaluation of the model predictions, both accuracy and F1-macro score are applied. Accuracy is one of the most intuitive performance evaluation metrics, but it may be misleading due to its sensitivity to class imbalance, which is why we also report the F1-macro score.

5 Experiment and Results

5.1 Experimental set-up

To evaluate the influence of sentiment analysis for short lines as a sub-task in the multi-task framework, we take the single task of sentiment analysis for the overall poems as the baseline.

We choose the dataset FSPC in our experiments. Considering the imbalanced distribution of the original five category labels, as shown in Figure 2, we merge the label *implicit negative* into *negative*, and *implicit positive* into *positive*. As the prediction difficulty of each poem may vary, ten-fold cross-validation (Kohavi, 1995) is applied in the experiment, providing a more reliable estimate of the model’s performance compared to a single train-test split. The results reported later are the average results of the ten folds.

We designed two experiments: experiment 1 compares the influence of the framework complexity on the model prediction, while experiment 2 compares the performance of frameworks based on different pre-trained models.

Experiment 1 In this experiment, two modes are designed with the pre-trained model SikuBERT, namely the multi-task mode and the single-task mode, with the latter including only the sentiment analysis task of the overall poem, used for the ablation study. For the multi-task mode, HAN is added and then additional information from the short line labels is combined with HAN to the network. It is expected that HAN is able to better capture the hierarchical structure of classical Chinese poetry.

Experiment 2 Apart from SikuBERT, we also investigated other pre-trained models, i.e. BERT-base-Chinese, SikuRoberta, BERT-ancient-Chinese and BERT_CCPoem within the best framework from Experiment 1.

Hyperparameters For a fair comparison, all experiments are performed with the same dataset and hyperparameter settings. The optimizer is AdamW, with a learning rate of 1e-6. The epoch is set to 10 in experiment 1 and 20 in experiment 2, while the batch size is always set to 8. The mechanism of early stopping is applied in the experiment and the monitor is set as evaluation loss while the patience is set to 3. It should be noted that in both experiment 1 and 2, the early stopping is always triggered before the training reaches the epoch limit.

5.2 Experimental results

Benchmark To the best of our knowledge, there are recently 2 published studies on 3-class sentiment prediction with FSPC, evaluated with accuracy or F1 scores, and the best model achieved an F1-macro score of 64.8%, as shown in Table 1. Zhang et al. (2022) proposed to extract word

Studies	Acc	F1
Zhang et al. (2022)	NA	64.38
Hong et al. (2023)	67.10	64.80

Table 1: Benchmark results on the sentiment analysis task trained and evaluated with the FSPC dataset only.

Method	Acc	F1
single-task	69.00	66.27
multi-task	69.32	66.50
multi-task + HAN	70.06	67.49
multi-task + HAN + line label	70.96	68.51

Table 2: Accuracy and F1-macro scores on the overall poem sentiment analysis task on the FSPC dataset based on the SikuBERT pre-trained model.

vectors with two pre-trained models separately and then fuse them to enrich text vector information, while Hong et al. (2023) built a hybrid corpus of classical Chinese poetry with its modern Chinese translation, on the basis of which they fine-tuned a ChineseBERT model (Sun et al., 2021). Both these studies experimented with overall poem sentiment only.

Experiment 1 As shown in Table 2, the sentiment analysis performance improves with the increase in model complexity. To start with, the accuracy and F1-macro of the single-task model are 69.00% and 66.27% respectively, which already outperforms previous models in Table 1, which suggests the advantage of pre-trained models specifically designed for ancient Chinese, as the studies of both Zhang et al. (2022) and Hong et al. (2023) are based on pre-trained models for modern Chinese only.

Compared with the single-task model, the multi-task model achieves slightly higher scores in both accuracy (+0.32%) and F1 (+0.23%), suggesting the positive influence of the additional short line information on the performance of the overall poem sentiment prediction. Moreover, when the HAN is applied to the framework, a more noticeable improvement is witnessed, with an increase of 0.74% in accuracy and 0.99% in macro F1, indicating the advantage of HAN in dealing with structured texts. Furthermore, the introduction of short line labels has a positive influence on the model performance, reaching an accuracy score of 70.96% and an F1-macro score of 68.51%. We could thus conclude that leveraging both the poem structure and the fine-grained sentiment information at the line level

Method	Acc	F1
BERT_CCPoem	67.54	65.24
BERT-base-Chinese	69.60	67.33
BERT-ancient-Chinese	70.28	68.31
SikuBERT	70.96	68.51
SikuRoBERTa	72.88	71.05

Table 3: Accuracy and F1-macro scores of the overall poem sentiment analysis task using different pre-trained models in combination with the best framework in Experiment 1 (multi-task+HAN+line label).

is advantageous for sentiment classification of the classical Chinese poems.

Experiment 2 Results of other pre-trained models with the best framework in Experiment 1 are shown in Table 3. The F1-macro scores range from 65.24% to 71.05%, the latter being obtained by SikuRoBERTa.

BERT-base-Chinese is the only model that is not pre-trained specifically for ancient Chinese. Considering the difference between ancient Chinese and modern Chinese, it is not surprising that BERT-base-Chinese does not perform better with classical Chinese poems than most of the other models, except the model BERT_CCPoem. Although BERT_CCPoem is the only model pre-trained specifically for Chinese ancient poems in the experiments, it shows worse accuracy and macro F1-scores compared to the other models. One possible reason could be that the training data of BERT_CCPoem is limited to poems, resulting in a smaller vocabulary of 11,809, while other models, such as BERT-base-Chinese and BERT-ancient-Chinese have a larger vocabulary of 21,128 and 38,208 respectively.

The SikuRoBERTa-based model within the proposed framework (multi-task+HAN+line label) in our experiments achieves the best performance, with an accuracy of 72.88% and an F1-macro score of 71.05%, outperforming previous studies, as shown in Table 1.

6 Prediction Analysis

With the ten-fold cross-validation in Experiment 2, we obtained predictions for all 5,000 poems using the SikuRoBERTa model in the multi-task framework with line-label-enhanced hierarchical attention, as shown in Figure 5.

First of all, it is clear that the model performs better with poems with clear sentiment polarities,

True.		Pred.			
		neg	neu	pos	
neg	1420 (80.91%)	254 (14.47%)	81 (4.62%)		100%
neu	194 (14.61%)	691 (52.03%)	443 (33.36%)		100%
pos	63 (3.29%)	321 (16.74%)	1553 (79.97%)		100%
		neg	neu	pos	

Figure 5: Sentiment prediction distribution of SikuRoBERTa-based multi-task framework with enhanced hierarchical attention. The X-axis represents the predicted labels, while the Y-axis stands for the true labels

i.e., negative and positive, reaching an accuracy of around 80%, while the model has only fair predictions with neutral poems, with an accuracy of about 52%. For the negative and positive poems, more wrong predictions are distributed on their nearest sentiment neighbour, i.e., neutral, rather than the far/opposite sentiment neighbours, i.e., positive and negative respectively, suggesting the model’s ability to differentiate between positive and negative poems.

As for the neutral poems, the model achieves an accuracy of 52.03%, which is far less than the performance with negative and positive poems, indicating that differentiating between neutral and implicit sentiment is still challenging. Although only about half are correctly predicted, around one-third of the wrong predictions are predicted as negative and the rest as positive, with the latter outnumbering and even doubling the former, indicating that the model tends to have more non-negative predictions on neutral poems. To gain more insights into this tendency, we further investigated the short line labels of 194 false negative poems and 443 false positive poems which are labelled as neutral by the annotators as shown in Figure 5. For both the false negative and false positive groups of poems, we plotted the sentiment label distribution of each line group (i.e. all first lines are joined into one group, the same holds for the second lines, etc.), and the results are shown in Figure 6 and 7.

As shown in Figure 6, for the 194 neutral poems that are predicted as negative by the model, line 1 and line 2 seem to hold sentiment labels that

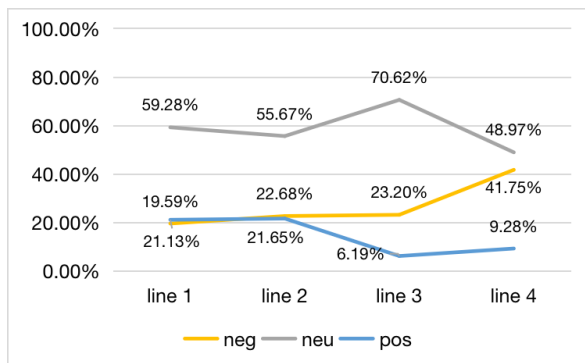


Figure 6: True sentiment label distribution across short lines in 194 neutral poems that are predicted as negative.

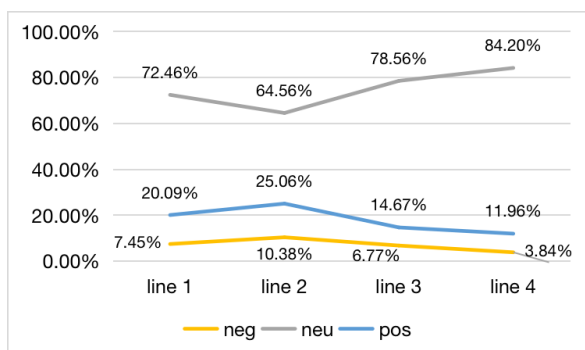


Figure 7: True sentiment label distribution across short lines in 443 neutral poems that are predicted as positive.

contain almost equal percentages of negative and positive labels. In line 3, however, there is more negative sentiment labelled than positive and this tendency continues in line 4, where the negative sentiment reaches 41.75%, which is more than four times of the number of positive annotations. On the other hand, as shown in Figure 7, when we consider the sentiment annotations at the line level of the 443 neutral poems that were predicted as positive by the model, we see an opposite tendency with always more positive sentiment labels than negative ones. Based on the sentiment distribution across short lines, we hypothesize that although the model fails to make correct predictions for about half of the neutral poems, it still senses the more implicit sentiment polarities in the poems, leading to negative predictions for the more negative poems and positive predictions for the more positive poems.

Moreover, we also investigate specific poems where there is a *significant difference* between the true labels and the predictions. By *significant difference*, we refer to the cases where the poem is classified with the complete opposite polarity of its labeled polarity, e.g. the overall sentiment of

Poem	True labels		Prediction	
	line	overall	line	overall
珠树森森秀阮林 The pearl trees flourish, lush as in the Ruan Forest	neutral	negative	positive	positive
高堂有母各欢心 In the grand hall, the mother's heart is filled with joy	positive		positive	
兰陔娱养时多暇 Amidst the fragrant garden, time for care is ample	neutral		positive	
勿遣平安阙嗣音 Let not the news of peace be left unheard by the heir	negative		neutral	

Figure 8: Comparison of the true labels and predictions of one poem. Note that the English translation comes from ChatGPT and works only as reference.

the poem is predicted as positive while the true label is negative, or vice versa. Of the 5,000 poems, there were in total 144 poems classified with this opposite polarity, and we found that this opposite polarity labelling happens both at the line level and the poem level. Figure 8 shows an example poem with its manual annotations and model predictions for the short lines and the overall poem. For the short lines, the model predicted positive instead of neutral for the first and the third lines, and predicted neutral instead of negative for the last line, generally upgrading the short line sentiment toward the polarity of positive, which might lead to the positive overall prediction. However, if contextual or topic information of the poem, which is “farewell”, would be provided, and if the model learns from other poems that “farewell” is usually related to negative emotions, the possibility of the model predicting this poem as overall negative would increase.

7 Conclusion

In this paper, we introduced a multi-task framework with enhanced hierarchical attention for sentiment analysis on classical Chinese poetry. This multi-task framework consists of two sub-tasks, the sub-task of sentiment analysis on the short lines in the poem, and the sub-task of sentiment analysis on the overall poem. For the latter task, a hierarchical attention network composed of word- and sentence-level attention was applied. Furthermore, to further utilize the information from short lines, additional information from short line sentiment was introduced to the sentence-level attention.

Experiments on the FSPC dataset show that our framework, compared with the single-task setup and based on the pre-trained model SikuBERT,

yields a increase of macro F1 from 66.27% to 68.51%. We also experimented with other pre-trained models, and the best performance is demonstrated by SikuRoBERTa, with an accuracy of 72.88% and an F1-macro of 71.05%, thus largely outperforming the state-of-the-art with an increase 5.78% in accuracy and 6.25% in F1-macro score (Hong et al., 2023).

We also investigated the SikuRoBERTa-based model predictions in more detail. We found that the model achieves an accuracy of about 80% with negative and positive poems, but only an accuracy of around 50% with neutral poems, which might be related to the “implicit emotion” writing style favoured in classical Chinese poems. Moreover, a further look at the predictions on the neutral poems suggested that the incorrect predictions are subtly aligned with the more prominent positive or negative sentiment polarities in the short lines, which indicates that regardless of the incorrect predictions on the neutral poems, the model still senses whether the poems are more positive or negative.

8 Future Work

The multi-task framework proposed in this paper has demonstrated potential in the task of sentiment analysis on classical Chinese poetry. It would also be interesting to introduce the task of emotion recognition in the framework, although this might require more annotations. Moreover, as the introduction of short line labels helps to improve the model performance, it is also promising to include additional information, such as the background or the topic of the poem to make more accurate predictions. We will also investigate how these different types of information can be optimally fused.

9 Limitation

In the experiment of this paper, the results are limited to the dataset FSPC which contains 5000 instances. A larger dataset would help to produce a stronger statement.

10 Acknowledgements

This research received funding from the Flemish Government under the Research Program Artificial Intelligence (174E07824). We would also like to thank the anonymous reviewers for their valuable and constructive feedback.

References

- Shakeel Ahmad, Muhammad Zubair Asghar, Fahad Mazaed Alotaibi, and Sherafzal Khan. 2020a. Classification of poetry text into the emotional states using deep learning technique. *IEEE Access*, 8:73865–73878.
- Shakeel Ahmad, Muhammad Zubair Asghar, Fahad Mazaed Alotaibi, and Sherafzal Khan. 2020b. Classification of poetry text into the emotional states using deep learning technique. *IEEE Access*, 8:73865–73878.
- Georgios Balikas, Simon Moura, and Massih-Reza Amini. 2017. Multitask learning for fine-grained twitter sentiment analysis. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1005–1008.
- Abdessamad Chanaa et al. 2021. E-learning text sentiment classification using hierarchical attention network (han). *International Journal of Emerging Technologies in Learning (iJET)*, 16(13):157–167.
- Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 281–290.
- Huimin Chen, Xiaoyuan Yi, Maosong Sun, Cheng Yang, Wenhao Li, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China.
- Jiajun Cheng, Shenglin Zhao, Jiani Zhang, Irwin King, Xin Zhang, and Hui Wang. 2017. Aspect-level sentiment classification with heat (hierarchical attention) network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 97–106.
- Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM international conference on data mining*, pages 235–243. SIAM.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Rodolfo Delmonte et al. 2013. Computing poetry style. In *Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive*

- Media: approaches and perspectives from AI (ESSEM 2013)*, volume 1096, pages 148–155. CEUR Workshop Proceedings (CEUR-WS.org).
- Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515.
- Jie Hong, Tingting He, Jie Mei, Ming Dong, Zheming Zhang, and Xinhui Tu. 2023. A hybrid corpus based fine-grained semantic alignment method for pre-trained language model of ancient Chinese poetry. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4794–4801. IEEE.
- Yufang Hou and Anette Frank. 2015. Analyzing sentiment in classical Chinese poetry. In *Proceedings of the 9th SIGHUM workshop on language Technology for Cultural Heritage, social sciences, and humanities (LaTeCH)*, pages 15–24.
- Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 8–17.
- Jasleen Kaur and Jatinderkumar R Saini. 2014. Emotion detection and sentiment analysis in text corpus: a differential study with informal and formal writing styles. *International Journal of Computer Application*, ISSN, pages 0975–8887.
- Jasleen Kaur and Jatinderkumar R Saini. 2020. Designing Punjabi poetry classifiers using machine learning and different textual features. *Int. Arab J. Inf. Technol.*, 17(1):38–44.
- Evgeny Kim and Roman Klinger. 2021. A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift für digitale Geisteswissenschaften*.
- Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95*, page 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Gang Li and Jie Li. 2018. Research on sentiment classification for Tang poetry based on TF-IDF and FPGrowth. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 630–634. IEEE.
- James JY Liu. 2022. *The art of Chinese poetry*. Routledge.
- Yutong Liu, Bin Wu, and Ting Bai. 2020. The construction and analysis of classical Chinese poetry knowledge graph. *Journal of Computer Research and Development*, 57(6):1252–1268.
- Adria Mallo-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. In *20th Annual Conference of the International Speech Communication Association (INTER-SPEECH 2019)*, pages 221–225.
- László Nemes and Attila Kiss. 2021. Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, 5(3):375–394.
- Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1015–1025.
- Nuttachot Promrit and Sajjaporn Waijanya. 2017. Convolutional neural networks for Thai poem classification. In *Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part I 14*, pages 449–456. Springer.
- Annie Rajan and Ambuja Salgaonkar. 2020. Sentiment analysis for Konkani Language: Konkani Poetry, a case study. In *ICT Systems and Sustainability: Proceedings of ICT4SD 2019, Volume 1*, pages 321–329. Springer.
- Somiya Rani and Amita Jain. 2024. Aspect-based sentiment analysis of drug reviews using multi-task learning based dual BiLSTM model. *Multimedia Tools and Applications*, 83(8):22473–22501.
- Yabo Shen, Yong Ma, Chunguo Li, Shidang Li, Mingliang Gu, Chaojin Zhang, Yun Jin, and Yingli Shen. 2019. Sentiment analysis for Tang poetry based on imagery aided and classifier fusion. In *Artificial Intelligence for Communications and Networks: First EAI International Conference, AICON 2019, Harbin, China, May 25–26, 2019, Proceedings, Part II*, pages 283–290. Springer.
- Apoorva Singh, Sriparna Saha, Md Hasanuzzaman, and Kuntal Dey. 2022. Multitask learning for complaint identification and sentiment analysis. *Cognitive Computation*, 14(1):212–227.
- Rachele Sprugnoli, Francesco Mambrini, Marco Passarotti, Giovanni Moretti, et al. 2022. Sentiment analysis of Latin poetry: First experiments on the Odes of Horace. *Computational Linguistics CliC-it 2021*, page 314.
- Chang Su, Shupin Liu, and Chalian Luo. 2023. MISC: A multimodal approach for sentiment classification of classical Chinese poetry. In *International Conference on Intelligent Computing*, pages 432–442. Springer.

- Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. **ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075, Online. Association for Computational Linguistics.
- Yik Yang Tan, Chee-Onn Chow, Jeevan Kanesan, Joon Huang Chuah, and YongLiang Lim. 2023. Sentiment analysis and sarcasm detection using deep multi-task learning. *Wireless personal communications*, 129(3):2213–2237.
- Yongrui Tang, Xumei Wang, Peng Qi, and Yan Sun. 2020. A neural network-based sentiment analysis scheme for Tang poetry. In *2020 International Wireless Communications and Mobile Computing (IWCWC)*, pages 1783–1788. IEEE.
- Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. AnchiBERT: A pre-trained model for ancient Chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, et al. 2023. GujiBERT and GujiGPT: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.
- Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen, and Bin Li. 2022. Sikubert and sikuroberta: Construction and application of pre-trained models of Siku Quanshu in orientation to digital humanities. *Library Tribune*, 42(6):31–43.
- Pengyu Wang and Zhichen Ren. 2022. The uncertainty-based retrieval framework for ancient Chinese CWS and POS. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168.
- Yuting Wei, Huazheng Wang, Jiaqi Zhao, Yutong Liu, Yun Zhang, and Bin Wu. 2020. GeLaiGeLai: a visual platform for analysis of classical Chinese poetry based on knowledge graph. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 513–520. IEEE.
- Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954.
- Zhaozhen Wu, Guoyi Cao, and Wanghao Mo. 2023. Multi-tasking for aspect-based sentiment analysis via constructing auxiliary self-supervision acop task. *IEEE Access*.
- Chen Xia. 2021. Poetry and emotion in classical Chinese literature. In *The Routledge Handbook of Chinese Studies*, pages 289–303. Routledge.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.
- Shanshan Yi and Xiaofang Liu. 2020. Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers’ review. *Complex & Intelligent Systems*, 6(3):621–634.
- Lingli Zhang, Yadong Wu, Qikai Chu, Pan Li, Guijuan Wang, Weihang Zhang, Yu Qiu, and Yi Li. 2022. **SA-model: multi-feature fusion poetic sentiment analysis based on a hybrid word vector model**. In *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 984–988.
- Wei Zhang, Hao Wang, Min Song, and Sanhong Deng. 2023. A method of constructing a fine-grained sentiment lexicon for the humanities computing of classical Chinese poetry. *Neural Computing and Applications*, 35(3):2325–2346.
- Guoshuai Zhao, Yiling Luo, Qiang Chen, and Xueming Qian. 2023. Aspect-based sentiment analysis via multitask learning for online reviews. *Knowledge-Based Systems*, 264:110326.
- Huidong Zhao, Bin Wu, Haoyu Wang, and Chuan Shi. 2014. Sentiment analysis based on transfer learning for Chinese ancient literature. In *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESCC2014)*, pages 1–7. IEEE.

Exploring Similarity Measures and Intertextuality in Vedic Sanskrit Literature

So Miyagawa

The University of Tsukuba
miyagawa.so.kb@u.tsukuba.ac.jp

Yuki Kyogoku

Leipzig University
kyogoku11@gmail.com

Yuzuki Tsukagoshi

The University of Tokyo
yuzuki@l.u-tokyo.ac.jp

Kyoko Amano

Kyoto University
amano.skskrt@kcn.jp

Abstract

This paper examines semantic similarity and intertextuality in selected texts from the Vedic Sanskrit corpus, specifically the Maitrāyaṇī Saṃhitā (MS) and Kāṭhaka Saṃhitā (KS). Three computational methods are employed: Word2Vec for word embeddings, stylo package for stylometric analysis, and TRACER for text reuse detection. By comparing various sections of the texts at different granularities, patterns of similarity and structural alignment are uncovered, providing insights into textual relationships and chronology. Word embeddings capture semantic similarities, while stylometric analysis reveals clusters and components that differentiate the texts. TRACER identifies parallel passages, indicating probable instances of text reuse. The computational analysis corroborates previous philological studies, suggesting a shared period of composition between MS.1.9 and MS.1.7. This research highlights the potential of computational methods in studying ancient Sanskrit literature, complementing traditional approaches. The agreement among the methods strengthens the validity of the findings, and the visualizations offer a nuanced understanding of textual connections. The study demonstrates that smaller chunk sizes are more effective for detecting intertextual parallels, showcasing the power of these techniques in unraveling the complexities of ancient texts.

1 Introduction

Vedic Sanskrit literature preserves invaluable cultural and historical information from ancient India. However, their study presents unique challenges due to linguistic characteristics, modes of composition, and transmission. Computational methods offer promising avenues to analyze such texts on an unprecedented scale. In this paper, we explore similarity measures and intertextuality between selected texts from the Vedic corpus - the Maitrāyaṇī Saṃhitā (MS) and Kāṭhaka Saṃhitā (KS). These

texts belong to different śākhās or schools, and are considered to exhibit mutual influence in their composition around 900–700 BCE.¹

The main focus of this paper is to present reliable numerical data on the chapter-wise similarity between the MS and KS. While it is known that the MS and KS are parallel texts, the variations in similarity among chapters have not yet been confirmed through numerical data. Since differences in chapter-wise similarity can contribute to estimating the relative chronology of each chapter, this similarity analysis holds significant importance for understanding the process of textual composition.

In recent years, the editorial process has been increasingly elucidated through philological studies (Amano, 2014-2015, 2020), suggesting variations in similarity between different sections in MS and KS depending on the time period. That is to say, sections edited in earlier periods exhibit lower similarity between MS and KS, whereas those edited in later times show higher similarity between MS and KS. Similarity analysis using computational methods further advances this study.

Our analysis employs three approaches:

1. Word embeddings generated using Word2Vec
2. Stylometry analysis using the stylo package
3. text reuse detection with TRACER

The word embeddings approach vectorizes the texts and compares the cosine similarity of the vectors. The stylo (Eder et al., 2016) and TRACER (Büchler, 2013; Büchler et al., 2018) approaches examine stylistic similarity and text reuse at document level.

The texts are pre-processed by undoing phonological change (sandhi) in the original texts and

¹All the corpora, codes, and results are available on our GitHub repo <https://github.com/somiyagawa/VedicSanskrit> (accessed October 5, 2024).

lemmatizing the words. Different chunking of the text is compared — at section level and by segments of 20, 100 and 200 words.

The results demonstrate interesting patterns of similarity and clustering between different text segments, with general alignment between the three approaches. This research highlights the potential of computational methods in studying ancient languages and aims to inspire further collaborative research at the intersection of Indology and computational linguistics.

2 Related Work

Computational methods have been increasingly applied to study various aspects of Sanskrit literature in recent years. [Hellwig et al. \(2020\)](#) developed a neural network architecture for processing Sanskrit texts. [Krishna et al. \(2019\)](#) analyzed poetic style in Sanskrit poetry using deep learning techniques.

Regarding Vedic Sanskrit specifically, [Hellwig et al. \(2023\)](#) developed a dependency parser for Ṛgvedic Sanskrit. [Hellwig and Nehrlich \(2018\)](#) compiled a Vedic treebank. These works provide NLP tools and resources for computational processing of Vedic texts.

Stylometry has been widely used to study authorship and stylistic similarity in classical literature. For instance, [Stover et al. \(2016\)](#) applied stylometric analysis using the *stylo* R package ([Eder et al., 2016](#)) to investigate the authenticity of an unknown classical Latin text called the *Expositio*. Their study concluded that this work was probably written by the second-century African author Apuleius of Madauros.

While stylometry focuses on authorship attribution and stylistic analysis on a macro-level, as demonstrated by the *stylo* package, text reuse detection tools offer a micro-level approach to detecting each text reuse such as quotations and allusions among texts. Specifically, TRACER ([Büchler, 2013](#); [Büchler et al., 2014](#); [Büchler et al., 2018](#)) is a text reuse detection tool that has been successfully applied to study intertextuality in ancient Greek ([Buechler et al., 2008](#); [Büchler et al., 2010](#)), Latin ([Franzini et al., 2018b](#)), Coptic texts ([Miyagawa, 2022, 2021](#); [Miyagawa et al., 2018](#)), Classical Tibetan ([Almogi et al., 2019](#)), German ([Franzini et al., 2018a](#)), etc.

Other programs are also available for historical text reuse analysis. For example, Tesseract and Passim are well-known tools in this field. Tesseract

([Coffee et al., 2012](#)) is primarily used for Latin texts, while Passim ([Romanello and Hengchen, 2021](#)) has been adapted for Western languages and Arabic with promising results but has not yet been adapted for Sanskrit. Compared to these tools, TRACER offers greater flexibility and customizability, making it possible to adapt it to Vedic Sanskrit using custom lemmatization, synonym, and cohyponym files.

3 Methodology

3.1 Corpus

The corpora consist of selected texts from the Maitrāyaṇī Saṃhitā (MS) and Kāṭhaka Saṃhitā (KS). The following sections are analyzed:

1. MS.1.1 (MS.1.1.1-1.1.13): 1145 words
2. MS.1.6 (MS.1.6.3-13): 3816 words
3. MS.1.7 (MS.1.7.2-5): 819 words
4. MS.1.9 (MS.1.9.3-8): 1627 words
5. KS.8 (KS.7.15 + 8.1-12): 3519 words
6. KS.9.1 (8.15 + 9.1-3): 818 words
7. KS.9.11 (9.11-17): 1721 words

The corpus MS.1.1 includes ritual formulas for new and full moon sacrifice. MS.1.6 includes ritual explanation about establishment of sacred fires, whose parallel is KS.8. MS.1.7 includes ritual explanation about reestablishment of sacred fires, whose parallel is KS.9.1. MS.1.9 includes explanation of secret spells related to ritualistic communal life, whose parallel is KS.9.11. MS.1.6, 1.7, and 1.9 have been philologically studied by [Amano \(2009\)](#), and their parallels in KS have been accurately identified. MS and KS were composed in the same editorial policy, and have almost the same contents for the same rituals, but with some variants in details. MS and KS contain portions with different linguistic styles and content (sometimes irregularly inserted or arranged), which necessitates the exclusion of such portions to conduct linguistically and semantically accurate analyses. The corpora used in this analysis were created to ensure that different styles (formulas or explanations) and contents (rituals) are not mixed. The chapter numbers within the parentheses following each section name represent the exact chapter numbers included in the

section. The size (word count) of each corpus is also provided above.

Using these corpora, we conduct comparisons (similarity analyses) between sections as follows:

1. MS.1.1 ↔ MS.1.6
2. MS.1.6 ↔ MS.1.7
3. MS.1.6 ↔ KS.8
4. MS.1.7 ↔ KS.9.1
5. MS.1.9 ↔ KS.9.11

The first two comparisons, namely MS.1.1 ↔ MS.1.6 and MS.1.6 ↔ MS.1.7, serve as an evaluation of the proposed methods, as their similar or dissimilar relations are philologically demonstrated. MS.1.1 differs significantly in content from MS.1.6 and MS.1.7, while the latter two share similar contents. Accordingly, if the proposed methods work well, the comparison of MS.1.1 ↔ MS.1.6 is expected to show a low similarity, whereas MS.1.6 ↔ MS.1.7 is expected to demonstrate a high similarity, compared to the former comparison.

The following three comparisons are between the texts of MS and KS. Each chapter is thought to have been edited in different periods and under different cultural influence, and therefore, the degree of similarity between MS and KS varies. The two comparisons, MS.1.6 ↔ KS.8 and MS.1.7 ↔ KS.9.1, were manually calculated (Amano, 2014-2015). As a result, the comparison MS.1.6 ↔ KS.8 showed a low similarity, while the comparison MS.1.7 ↔ KS.9.1 showed a very high similarity. Since it is philologically inferred that MS.1.6 is older than MS.1.7, the paper presented the perspective that chapters compiled in the earlier period have lower similarity with KS, whereas those from a later period have higher similarity with KS, indicating possible intertextual borrowing. From this, in our current analysis, the comparison of MS.1.6 ↔ KS.8 is anticipated to reveal a low similarity, while MS.1.7 ↔ KS.9.1 is expected to exhibit a high similarity.

In contrast, the last comparison, MS.1.9 ↔ KS.9.11, was not examined in the previous studies, and serves as the main focus of our current analysis, aiming to demonstrate to what extent this comparison shows similarity. If the comparison of MS.1.9 ↔ KS.9.11 reveals a high similarity, akin to MS.1.7 ↔ KS.9.1, it strongly suggests that the intertextual contact between MS.1.9 and KS.9.11

occurred during a later period, characterized by a tendency for MS and KS to exhibit similarities, as argued in Amano (2020).

The texts are procured from the Digital Corpus of Sanskrit.² Original Sanskrit text undergoes phonetic fusion and changes at word boundaries, known as sandhi. These fusions and changes make it challenging to segment the text into individual words and perform morphological analysis. Therefore, as a first step in processing the text, it is necessary to resolve the sandhi to create an "un-sandhi-ed" text, which can then be used for lemmatization. The texts stored in the Digital Corpus of Sanskrit are processed into un-sandhi-ed texts as well as lemmatized texts by the computational method of (Hellwig et al., 2020), verified through expert review and correction. For the purpose of comparing similarity, the lemmatized texts are used, which are manually divided into distinct chunks or paragraphs with attention to meaningful coherence:

- Section level
- Fixed-size segments of 20, 100 and 200 lemmas

3.2 Word Embeddings

Word embedding models capture semantic relationships between words from their co-occurrence in a large corpus. We use Word2Vec (Mikolov et al., 2013), a two-layer neural network that predicts surrounding context words given an input word. We employ the skip-gram library with the training algorithm set to the skip-gram and default parameters for other settings.

The training data consists of a collection of Vedic Sanskrit texts, excluding the MS and KS. The word embeddings are averaged for each segment to obtain a document vector. The similarity between document vectors is computed using cosine similarity. Cosine similarity is used to compare the document vectors by calculating the cosine of the angle between them. This measures how close the vectors are to each other while disregarding their magnitude.

3.3 Stylometry and Text Reuse

The stylometry analysis is performed using the stylo package in R (Eder et al., 2016). It supports a variety of statistical analyses to examine

²<http://www.sanskrit-linguistics.org/dcs/index.php> (accessed May 24, 2024)

stylistic similarity between texts, such as cluster analysis, multidimensional scaling, principal component analysis etc. We use the cosine similarity as the similarity metric. For text reuse detection, we use TRACER (Büchler, 2013; Büchler et al., 2014; Büchler et al., 2018), which has been successfully applied to study intertextuality in various ancient language corpora. It provides a Java implementation to detect different types of text reuse such as quotations, allusions and idioms.

4 Results

4.1 Word Embedding

Table 1 shows the average cosine similarity between text segments using Word2Vec. In general, the similarity scores increase as the chunk size increases from 20 to 200 lemmas.

Table 1: Average cosine similarity using Word2Vec

Text Pair	Chunk Size		
	20	100	200
MS.1.1 ↔ MS.1.6	0.813	0.899	0.925
MS.1.6 ↔ MS.1.7	0.856	0.934	0.959
MS.1.6 ↔ KS.8	0.863	0.941	0.964
MS.1.7 ↔ KS.9.1	0.860	0.940	0.971
MS.1.9 ↔ KS.9.11	0.844	0.933	0.959

The comparisons of MS.1.1 ↔ MS.1.6 exhibits a lower similarity than that of MS.1.6 ↔ MS.1.7, suggesting effective performance of the analysis. However, the high similarity of MS.1.6 ↔ KS.8, which were expected to less similar, contradicts the previous findings. This discrepancy from the expectation may arise from the larger number of dissimilar chunks compared to similar ones, despite the existence of parallels between MS.1.6 and KS.8. The dissimilarity is highlighted by averaging the similarity values, because even within sections that are considered to have high similarity, segments that do not correspond exhibit low similarity, and such segments outside the parallel parts overwhelmingly outnumber the parallel ones. Instead of averaging the similarity values, the similarity between the two documents can be also assessed by their structural alignment and the similarity of their parallel segments, visualized using graphs such as heatmaps and histograms (Figures 1, 2, and 3). Heatmaps, particularly those based on 20 lemmas, provide the most accurate depiction of similarity between the chunks in parallel form.

In the heatmap of MS.1.7 and KS.9.1 (Figure 2), the diagonal line highlighted in a light color indicates a high similarity of the chunks, illustrating that these two sections share parallels in the same order. Conversely, the heatmap of MS.1.6 and KS.8 (Figure 1) does not exhibit such close parallelism. The heatmap of MS.1.9 and KS.9.11 (Figure 3) shows a similar pattern to that of MS.1.7 and KS.9.1. The histograms corroborate these findings: MS.1.6 and KS.8 contain few sentences with a high similarity above 0.95, while MS.1.7 with KS.9.1 and MS.1.9 with KS.9.11 do. This suggests that MS.1.7 and MS.1.9 were composed under similar conditions, in close contact with KS, likely during a later period of composition.

4.2 Stylometry

The stylo package is used to perform cluster analysis and principal component analysis (PCA) on the texts divided into 20-lemma and 100-lemma chunks.

Figures 4 and 5 show the resulting dendrograms. The cluster analysis results align with our expectations for the known evaluation comparisons. MS.1.1 is consistently separated from the other texts, confirming its distinct nature. The pairs of parallel sections in MS and KS are correctly grouped together, indicating their stylistic similarity. Importantly, the PCA results (Figure 6) provide insights into our main focus, the comparison of MS.1.9 ↔ KS.9.11. This pair shows a closer stylistic relationship compared to MS.1.6 ↔ KS.8, but similar to MS.1.7 ↔ KS.9.1. This suggests that MS.1.9 and KS.9.11 likely share a similar compositional context or period with MS.1.7 and KS.9.1, supporting our hypothesis of their later period of composition and closer intertextual relationship. These stylometric results, particularly the PCA, complement our findings from word embeddings and text reuse detection, providing a multi-faceted view of the textual relationships in our corpus.

4.3 Text Reuse

The text reuse detection using TRACER yields the following number of parallels between the text pairs in Table 2.

Detection of MS.1.1 ↔ MS.1.6 reveals no reuse (parallel sentence), and those of other sections show a number of reuse, which indicates the analysis functions appropriately. The detection of 100-lemma corpora provides the number of close parallels. The highest number of parallels are found be-

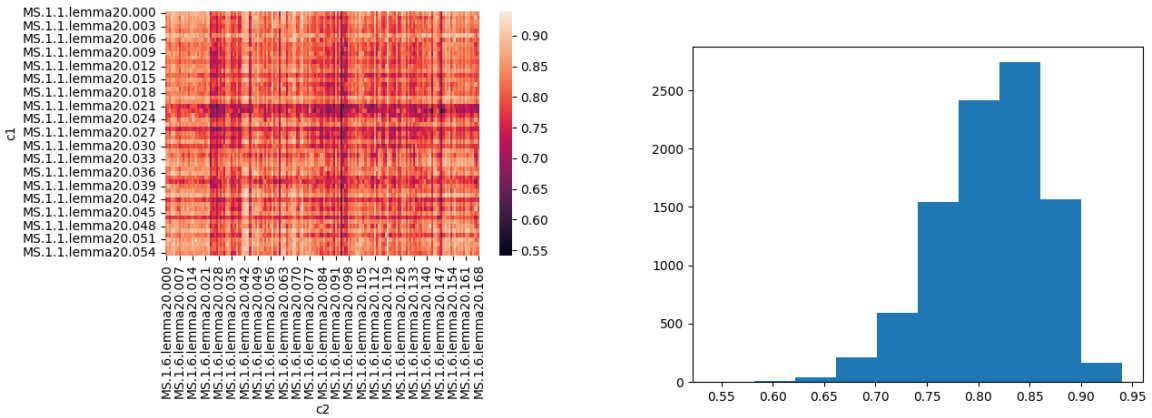


Figure 1: Word2Vec: heatmap and histogram of MS.1.6 ↔ KS.8 (20 lemma)

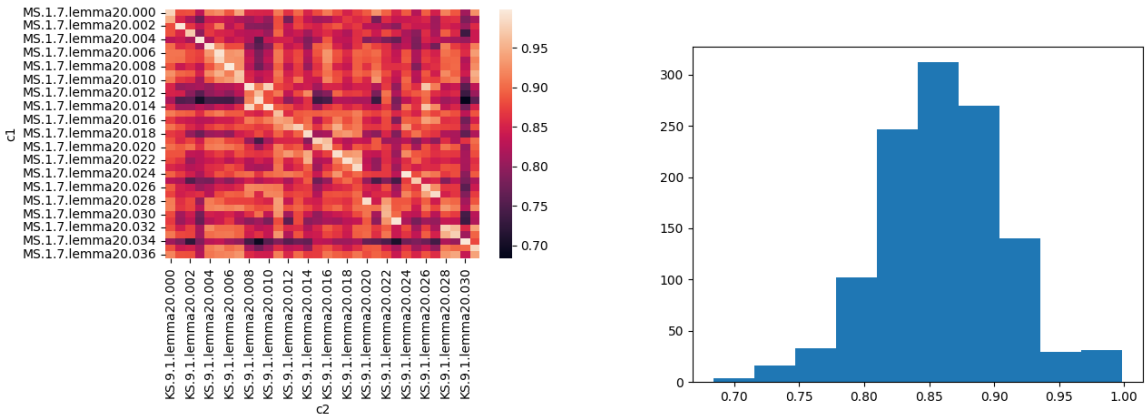


Figure 2: Word2Vec: heatmap and histogram of MS.1.7 ↔ KS.9.1 (20 lemma)

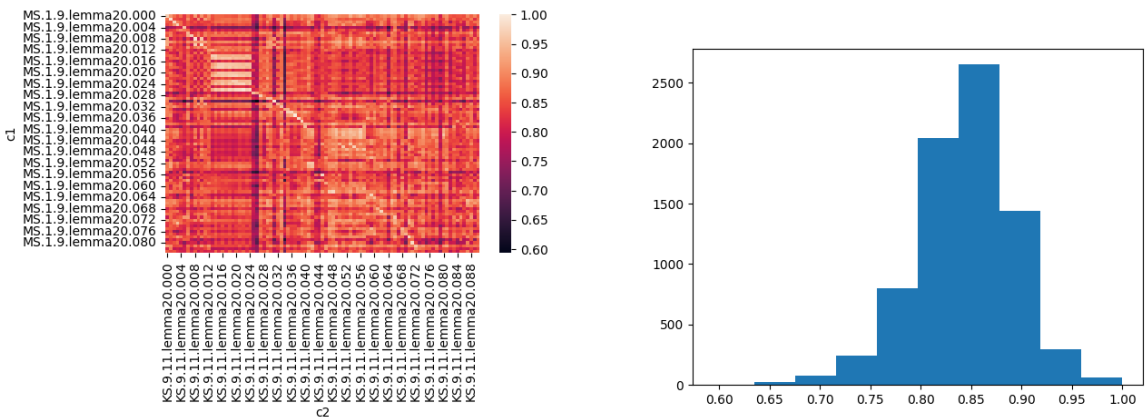


Figure 3: Word2Vec: heatmap and histogram of MS.1.9 ↔ KS.9.11 (20 lemma)

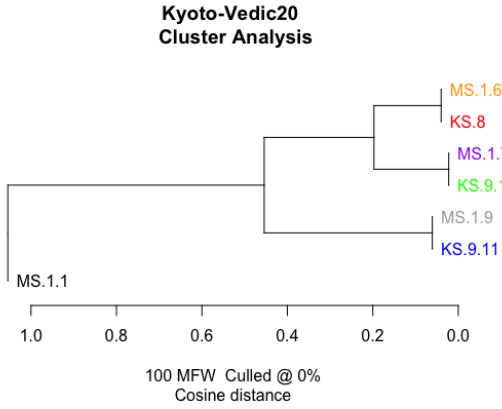


Figure 4: Cluster analysis of 20-lemma chunks using stylo

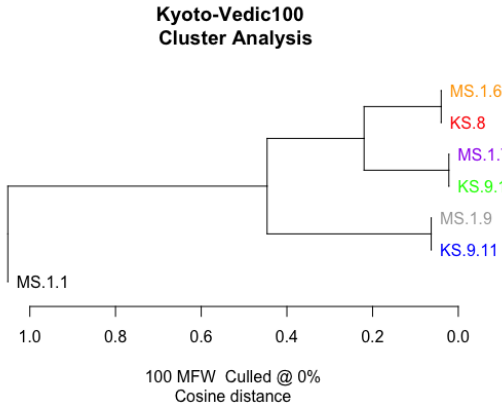


Figure 5: Cluster analysis of 100-lemma chunks using stylo

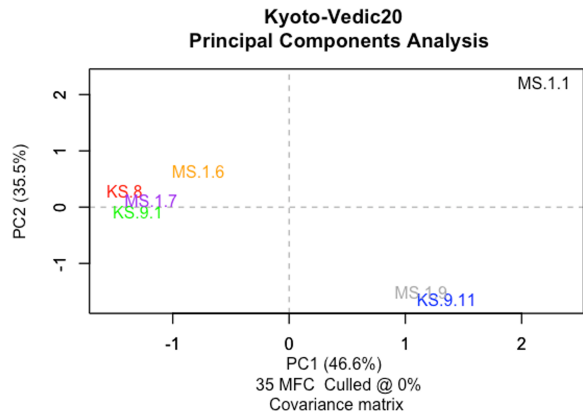


Figure 6: Stylo: Principal Components Analysis (20 lemma)

Table 2: Number of text reuse candidates detected by TRACER

Text Pair	20-lemma	100-lemma
MS.1.1 ↔ MS.1.6	N/A	N/A
MS.1.6 ↔ MS.1.7	13	3
MS.1.6 ↔ KS.8	8	15
MS.1.7 ↔ KS.9.1	55	10
MS.1.9 ↔ KS.9.11	209	15

tween MS.1.9 ↔ KS.9.11, followed by MS.1.7 ↔ KS.9.1. The detection of MS.1.6 ↔ KS.8 in 100-lemma corpus shows similar number of parallels to these two comparisons, which contradicts the previous study and of the analyses with Word2Vec and Stylo. The reason is that the size of the corpora is different (MS.1.6 contains 3816 words, MS.1.7 contains 819 words, MS.1.9 contains 1627 words). Due to the different sizes of the corpora, it is not appropriate to determine the similarity between sections based on the absolute number of parallels. However, graphs can compensate for this limitation.

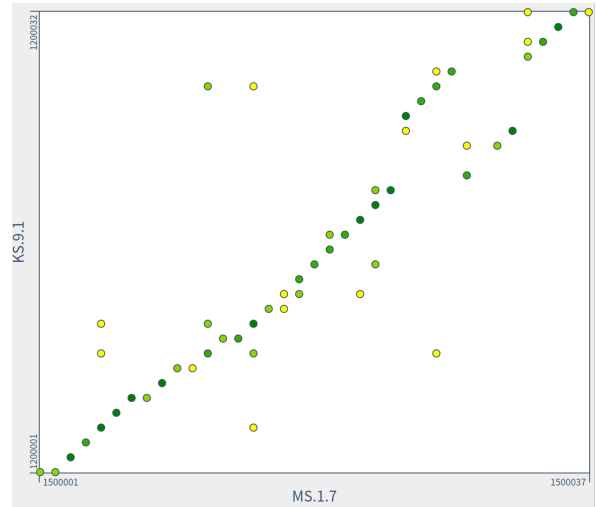


Figure 7: TRACER: MS.1.7 ↔ KS.9.1 (20 lemmas)

The graphs of MS.1.7 ↔ KS.9.1 (Figure 7) and MS.1.9 ↔ KS.9.11 (Figure 8) indicate the structural alignment, which is observed in the diagonal line of parallels, while that of MS.1.6 ↔ KS.8 (Figure 9) does not, as the heatmaps of Word2Vec indicated.

5 Conclusion

This paper presented an analysis of semantic similarity and text reuse in selected Vedic Sanskrit texts using word embedding, stylometric method and

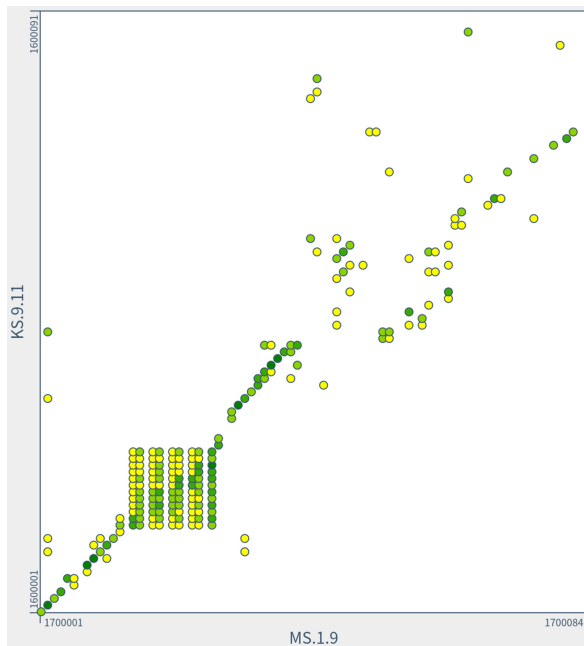


Figure 8: TRACER: MS.1.9 ↔ KS.9.11 (20 lemmas)

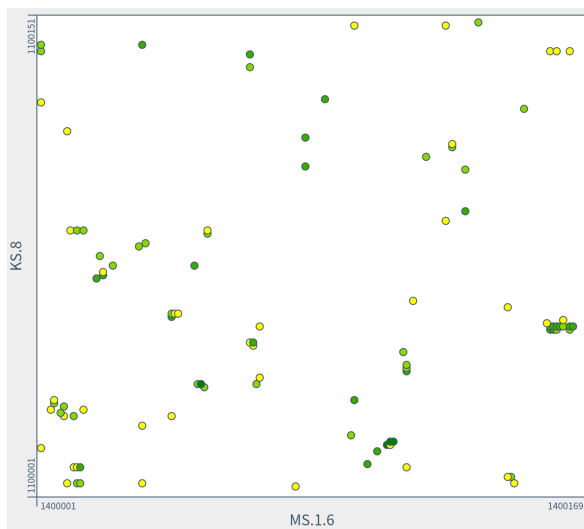


Figure 9: TRACER: MS.1.6 ↔ KS.8 (20 lemmas)

TRACER. The results from these approaches indicate patterns of similarity and clustering between different portions of the texts, which can be justified by previous literary studies. By analyzing the similarity of MS.1.9 ↔ KS.9.11, we inferred that MS.1.9 might share a similar historical period with MS.1.7.

- Using word embedding, the similarity between the pairs of sections was appropriately analyzed, the structural alignment was demonstrated well in the form of heatmap. The histograms helped us with understanding of similarity.
- The cluster analysis using stylo groups the corpora into intuitive clusters, with clearer separation at 100-lemma chunks.
- The text reuse detection using TRACER finds the highest number of parallels between MS.1.9↔KS.9.11 and MS.1.7↔KS.9.1, aligning with the stylometric clusters and the scores by the word embedding. The graphs show the structural alignment very well.

In conclusion, the computational analysis provides insights into the relationships between the texts and their sections in Vedic literature, which can clarify the process of its composition. The general agreement between the word embedding, text reuse detection and stylometric approach enhances the validity of the findings, and various visualizations of various analyses complemented each other's weaknesses and contributed to a more accurate understanding. Moreover, this study demonstrates that smaller chunk sizes are beneficial for finding parallels. On the other hand, documents with larger chunk sizes encompass various common topics. This means that documents with larger chunk sizes proportionally contain fewer topics that semantically distinguish them from each other, making it difficult to identify parallel relations with larger chunk sizes. Therefore, a smaller chunk size is more suitable for our purpose of finding parallels or verifying parallel relations between texts.

This research demonstrates the potential of computational methods in Vedic Sanskrit studies, and other ancient language corpora. Future work can extend the analysis to more texts, explore other embedding models and stylometric techniques, and closely examine the nature of the parallels identified. We hope this encourages further collaborative

research at the intersection of indology and computational linguistics.

Limitations

While our study provides valuable insights into the similarity and intertextuality in Vedic Sanskrit literature using computational methods, it is important to acknowledge certain limitations. The analysis is based on a limited corpus size, focusing on selected sections from two Vedic texts, and the pre-processing of the texts relies on one possible interpretation, which could lead to variations in the results. The Word2Vec model used may not fully capture the semantic nuances and complexities of Vedic Sanskrit, and more advanced models like BERT were not explored due to the limited size of the training dataset. The chunk sizes used for analysis were chosen based on meaningful coherence, but different sizes may provide additional insights. The stylometric analysis focused primarily on cluster analysis and principal component analysis, while other techniques could reveal further stylistic patterns. The text reuse detection effectively identifies parallel passages, but their significance requires further qualitative analysis by domain experts. It is important to note that the computational methods used are complementary to traditional philological and linguistic analysis, and integration with existing studies is crucial for a holistic understanding. Despite these limitations, our research demonstrates the potential of computational approaches in studying ancient languages and texts, and further interdisciplinary collaborations and advancements in computational methods can greatly contribute to this field of study.

Ethics Statement

This research aims to advance the understanding of ancient Vedic Sanskrit texts through computational methods while adhering to ethical considerations. The computational analysis complements traditional approaches, and the interpretation of results requires the expertise of Indologists and Sanskrit scholars. We recognize the cultural and religious significance of the Vedic texts and approach the analysis with respect and sensitivity. The methods and tools used are open-source, promoting transparency and reproducibility. We acknowledge the risk of misinterpretation or oversimplification and emphasize the need for caution in drawing conclusions. This research has the potential to

contribute to the preservation and understanding of ancient Indian heritage, inspiring further interdisciplinary research and public engagement. We are committed to conducting this research with integrity, transparency, and respect for the texts and the communities that hold them sacred.

References

- Orna Almogi, Lena Dankin, Nachum Dershowitz, and Lior Wolf. 2019. A hackathon for classical Tibetan. *Journal of Data Mining & Digital Humanities*, (Towards a Digital Ecosystem: NLP. Corpus infrastructure. Methods for Retrieving Texts and Computing Text Similarities).
- Kyoko Amano. 2009. *Maitrāyaṇī Saṃhitā I-II. Übersetzung der Prosapartien mit Kommentar zur Lexik und Syntax der älteren vedischen Prosa*, volume 9 of *Münchener Forschungen zur historischen Sprachwissenschaft*. Hempen Verlag, Bremen.
- Kyoko Amano. 2014-2015. Zur Klärung der Sprachschichten in der Maitrāyaṇī Saṃhitā. *Journal of Indological Studies*, 26/27:1–36.
- Kyoko Amano. 2020. What is ‘knowledge’ justifying a ritual action? uses of *ya evaṃ veda / ya evaṃ vidvān* in the Maitrāyaṇī Saṃhitā. In C. Redard, J. Ferrer-Losilla, H. Moein, and P. Swennen, editors, *Aux sources des liturgies indo-iraniennes*, volume 10 of *Collection Religions, Comparatisme - Histoire - Anthropologie*, pages 39–68. Presses Universitaires de Liège, Liège.
- Marco Büchler. 2013. Informationstechnische Aspekte des Historical Text Re-use.
- Marco Büchler, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. Towards a historical text re-use detection. In Michael W. Berry and Jacob Kogan, editors, *Text Mining*, pages 221–238. Springer.
- Marco Büchler, Greta Franzini, Emily Franzini, Maria Moritz, and Kirill Bulert. 2018. TRACER-a multi-level framework for historical text reuse detection.
- Marco Büchler, Annette Geßner, Gerhard Heyer, and Thomas Eckart. 2010. Detection of citations and textual reuse on ancient greek texts and its applications in the classical studies: eAQUA project. In *Proceedings of Digital Humanities 2010*, pages 113–114.
- Marco Buechler, Gerhard Heyer, and Sabine Gründer. 2008. eAQUA— bringing modern text mining approaches to two thousand years old ancient texts. In *Proceedings of e-Humanities—An Emerging Discipline, workshop at the 4th IEEE International Conference on e-Science*.
- Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W Forstall, Roelant Ossewaarde, and

- Sarah L Jacobson. 2012. The tesserae project: intertextual analysis of latin poetry. *Literary and linguistic computing*, 28(2):221–228.
- Maciej Eder, Jan Rybicki, and Mike Kestemont. 2016. [Stylometry with R: A Package for Computational Text Analysis](#). *The R Journal*, 8(1):107–121.
- Greta Franzini, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K Ochab, Emily Franzini, Joanna Byszuk, and Jan Rybicki. 2018a. Attributing authorship in the noisy digitized correspondence of jacob and wilhelm grimm. *Frontiers in Digital Humanities*, 5:4.
- Greta Franzini, Marco Passarotti, Maria Moritz, and Marco Büchler. 2018b. [Using and evaluating TRACER for an index fontium computatus of the summa contra gentiles of thomas aquinas](#). In Alessandro Mazzei Elena Cabrio and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it) 2018: Torino, Italy, December 10–12*, pages 199–205.
- Oliver Hellwig and Sebastian Nehrdich. 2018. [Sanskrit word segmentation using character-level recurrent and convolutional neural networks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium. Association for Computational Linguistics.
- Oliver Hellwig, Sebastian Nehrdich, and Sven Sellmer. 2023. Data-driven dependency parsing of vedic sanskrit. *Language Resources and Evaluation*, 57(3):1173–1206.
- Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. [The treebank of vedic Sanskrit](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France. European Language Resources Association.
- Amrith Krishna, Vishnu Sharma, Bishal Santra, Aishik Chakraborty, Pavankumar Satuluri, and Pawan Goyal. 2019. [Poetry to prose conversion in Sanskrit as a linearisation task: A case for low-resource languages](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1160–1166, Florence, Italy. Association for Computational Linguistics.
- Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). In *International Conference on Learning Representations*.
- So Miyagawa. 2021. Digitization of Coptic manuscripts and digital humanities: Tools and methods for Coptic studies. *The International Journal of Levant Studies*, 2:29–61.
- So Miyagawa. 2022. *Shenoute, Besa and the Bible Digital Text Reuse Analysis of Selected Monastic Writings from Egypt*. SUB Göttingen.
- So Miyagawa, Amir Zeldes, Marco Büchler, Heike Behlmer, and Troy Griffiths. 2018. Building linguistically and intertextually tagged coptic corpora with open source tools. In Chikahiko Suzuki, editor, *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*, pages 139–41. Center for Open Data in the Humanities.
- Matteo Romanello and Simon Hengchen. 2021. [Detecting text reuse with passim](#). *Programming Historian*, 10.
- Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology*, 67(1):239–242.

Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction

Laura Manrique-Gómez¹ Tony Montes² Arturo Rodríguez-Herrera³ Rubén Manrique²

¹ History and Geography Department, Universidad de los Andes, Bogotá D.C.

² Systems and Computing Engineering Department, Universidad de los Andes, Bogotá D.C.

³ Civil and Environmental Engineering Department, Rice University, Houston TX

{l.manriqueg, t.montes, rf.manrique}@uniandes.edu.co
da.rodriguez@rice.edu

Abstract

This paper presents two significant contributions: First, it introduces a novel dataset of 19th-century Latin American newspaper texts, addressing a critical gap in specialized corpora for historical and linguistic analysis in this region. Second, it develops a flexible framework that utilizes a Large Language Model for OCR error correction and linguistic surface form detection in digitized corpora. This semi-automated framework is adaptable to various contexts and datasets and is applied to the newly created dataset.

1 Introduction

The computational processing of historical newspaper texts is crucial due to the valuable information these texts contain about political, economic, and cultural history. Over the past three decades, Digital Humanities has driven extensive digitization efforts, resulting in numerous curated digital collections (Berry and Fagerjord, 2017; Dobson, 2019). However, converting these images into machine-readable texts remains challenging, particularly in achieving accurate transcription. A primary challenge is the accuracy of OCR technology, especially with the extremely diverse newspaper layouts, materially degraded documents, and non-standardized fonts typical of historical texts. Traditional OCR methods often produce errors that complicate subsequent analysis.

To address these challenges, we employed GPT-4o-mini (OpenAI, 2024), a Large Language Model (LLM), within a pipeline for OCR error correction. While the LLM is capable of fixing OCR-related errors that traditional systems often miss (Langlais, 2024), our pipeline also detects and classifies potential hallucinations to avoid further issues and streamline the process. Additionally, it contributes by identifying surface forms—specific word occurrences—within the dataset.

1.1 Related Work

The "Chronicling America" initiative marks a significant advancement in the digitization of historical newspaper materials (Humanities). Another major effort, is the "Atlas - Oceanic Exchanges" collection, which traces global information networks in 19th-century newspaper materials (Exchanges). Similarly, "Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines" (Cordell and Smith) explores the culture of reprinting in the U.S. before the Civil War, while the European "Project Impresso: Media Monitoring the Past" (SNSF and FNR, 2023) addresses the OCR challenges specific to English and Germanic languages.

Despite these advancements, historical newspapers are scarcely digitized in the Global South (LeBlanc, 2024). Consequently, a gap remains in specialized corpora for 19th-century Latin American newspapers, limiting the study of the region's unique historical and linguistic features. Our research addresses this gap by introducing a new dataset of Latin American newspaper texts in old Spanish. This dataset was post-processed with LLM models for addressing OCR errors and distinguishing them from historical linguistic surface forms¹.

ICDAR post-OCR correction competitions in 2017 and 2019 (Chiron et al., 2017; Rigaud et al., 2019) presented interesting solutions to error detection and correction in 10 European languages, such as Clova AI model based on multi-lingual BERT. Similarly, Nguyen et al. (2020) achieved comparable results by initializing embeddings with popular static embeddings such as GloVe (Pennington et al., 2014). In another approach, Veninga (2024) examined the fine-tuning of ByT5, a character-

¹The dataset is available at <https://huggingface.co/datasets/Flaglab/latam-xix> in its three versions: "original", "cleaned", and "corrected"

level LLM, emphasizing the importance of pre-processing and context length optimization. This results aligns with earlier studies on character-level models, such as Amrhein and Clematide (2018), which demonstrated the potential of character-based sequence-to-sequence models in improving OCR correction.

The application of LLMs for post-OCR correction has gained traction, especially in improving the accuracy of digitized historical texts. Early work by Nguyen et al. (2021) laid the foundation by categorizing post-OCR correction methods, highlighting the challenges associated with isolated-word and context-dependent approaches. As discussed by Thomas et al. (2024), the introduction of Transformers’ architecture leads to state-of-the-art performance in various text correction tasks and also presents a new baseline for post-OCR correction.

Langlais (2024) builds on this foundation by addressing the persistent issue of OCR quality in cultural heritage texts. They propose that LLMs can significantly enhance correction accuracy through context-aware processing, although challenges like hallucinations and language switching remain. More recent work by Thomas et al. (2024) demonstrates the superiority of a prompt-based approach using Llama 2 over traditional models like BART (Soper et al., 2021a), reducing character error rates (CER) by over 54%. These findings are consistent with those of Soper et al. (2021b), who reported comparable improvements using fine-tuned BART models. These studies highlight the evolution from traditional correction methods to LLM-based approaches. Nevertheless, further studies are needed to test correction methods in historical documents containing linguistic and regional variants.

2 Sourcing

The dataset was initially compiled from Colombian digital newspaper archives. The primary focus was on publications that included cartoons or illustrations, which were intended for subsequent multimodal modeling. This review also extended to the physical collections on-site, as only approximately 50% of the physical collection had been digitized. Through this process, 64 newspaper titles were identified, representing 7% of the total 1,655 publications in the collections. This first iteration resulted in a dataset consisting of 4,032 pages of scanned pages of newspapers, primarily from



Figure 1: El Oso, Peru. An example of a scanned newspaper image. The corresponding OCR-extracted text and the corrected version can be found in Appendix A, for reference.

Nueva Granada—a former country encompassing Colombia, Panama, Venezuela, and Ecuador—.

A second iteration completed the revision of 3,038 digitized newspapers of 58 digital collections across Mexico, Argentina, Colombia, Peru, Chile, Panama, Venezuela, Uruguay, Bolivia, Cuba, and Ecuador as shown in Table C1. Some countries, such as Bolivia, Cuba, and Venezuela have very limited or no web collections, resulting in their underrepresentation or absence from the final dataset. Additionally, some newspapers were printed in Europe due to lower costs; in some cases, printing outsourcing was utilized. The final dataset comprises 197 newspaper titles and 23,522 pages of scanned images, primarily from Mexico City (Mexico is the only country that has digitized its entire collection), but also includes publications from other Latin American cities, such as Buenos Aires, Lima, Bogota, and Santiago de Chile. An example of a newspaper image can be observed in Figure 1.

Originally, the Latin American 19-century newspaper dataset consists of scanned images. These images were processed using a layout model, followed by an OCR service. The layout model was specifically trained using data from annotated newspapers available in Roboflow OCR (2022); Alpha (2023); RSCOE (2023); GrabadosXIX (2023). These datasets were merged into a single dataset (CD) consisting of 1368 images of newspapers annotated for binary layout classification: images and texts. The CD dataset includes 10% of images from our newspapers dataset, labeled by hand, and it was enriched with data augmentation for shear and rotation. These techniques help to increase the model’s performance in images with scanning errors.

The CD dataset was used to train an image

recognition model from Azure Cognitive Services², which can extract the images in the newspaper page and extract the text through the OCR. The model's performance scored MAP@75 of 87.0%, resulting in a collection of annotations and coordinates for both text and images. These coordinates were used to crop the original image, and then process it with the OCR model. Once the OCR results were obtained, we merged the processed text with the images, creating a dataset that contains the newspaper images and their associated text. From a sample of 2,500 transcribed texts, each containing 1,000 characters, manual supervision revealed that 8.5% were unreadable. The remaining texts contained multiple transcription errors, primarily due to the artisanal printing techniques and the grammatical and lexical variations of the era. These errors significantly impacted readability, introducing bias when using the texts as input for NLP-LLM models.

3 Processing

The dataset includes samples of newspapers that were either handwritten or produced using early carving machines. Over time, these machines would wear out, leading to text features that were easily confused with backward accent marks, unwanted punctuations, or misplaced characters between words. Such misreadings disrupted the continuity of the text without adding any semantic meaning.

Detecting these errors automatically poses a challenge due to the linguistic shifts between modern and 19th-century Spanish. OCR models trained on such historical texts are lacking, especially considering the semantic and orthographic changes over time. For instance, what might appear as an OCR error could instead be a historical surface form of a word; for example, the conjunction "y" (and) was often written as "I".

Additionally, some texts were completely unintelligible for OCR, and challenging for humans to interpret, due to the fonts used in certain newspapers. The varied layouts of these newspapers also resulted in texts filled with scores or numbers, or in some cases, samples containing only chapter titles or numbering (e.g., "III IV V"), which added noise to the dataset. A general overview of the pipeline from the source until the final post-processed, is observed in Figure 2.

²Model available through Azure cloud services at <https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/>

3.1 Cleaning and filtering

Some of the most common cleaning steps for text data include removing duplicates and noisy data, which are particularly crucial for subsequent analysis. In this case, 3.08% of rows were removed due to duplicates or empty texts. Additionally, 1.74% of rows were filtered out where over 50% of the characters were non-alphabetic, as these rows are more likely to be noise than useful content. Rows with four or fewer tokens were also removed, accounting for 0.61% of the data; this was achieved by training a new tokenizer with a vocabulary size of 52,000, derived from the BETO (Spanish BERT) pre-trained tokenizer (Cañete et al., 2020).

3.2 Post-OCR LLM Correction

As previously discussed, LLMs have established a baseline for correcting OCR errors in historical texts (Thomas et al., 2024; Langlais, 2024). Detecting and fixing OCR errors from newspapers is challenging because these errors are often subtle and numerous. This problem is especially pronounced with 19th-century newspapers, where the quality of the paper and the outdated printing methods contribute to a high frequency of errors. These errors create significant noise and complicate the text correction process (Lopresti, 2008).

In this paper, we use a technique for detecting OCR errors and correcting them using GPT-4o-mini and taking advantage of the fact that LLMs were trained mostly in modern language. This way, manually checked rules can classify corrections between errors, word surface forms, or none of both (hallucinations). These rules, explained in the following section, were revised and selected by a field expert who served as well as an evaluator for these corrections testing their precision for this case.

We employed a *diff* algorithm to detect the differences between the original and corrected texts. This approach allowed us to fully leverage the LLM's ability to correct the text while ensuring a reliable and structured output. The *diff* algorithm identifies added, removed, and changed parts between the two texts, similar to the functionality seen in GitHub's blame feature. By doing so, we can specify the exact changes made during the correction process, enabling us to classify these alterations effectively.

This method proved more effective than instructing the LLM to return corrections in a specific

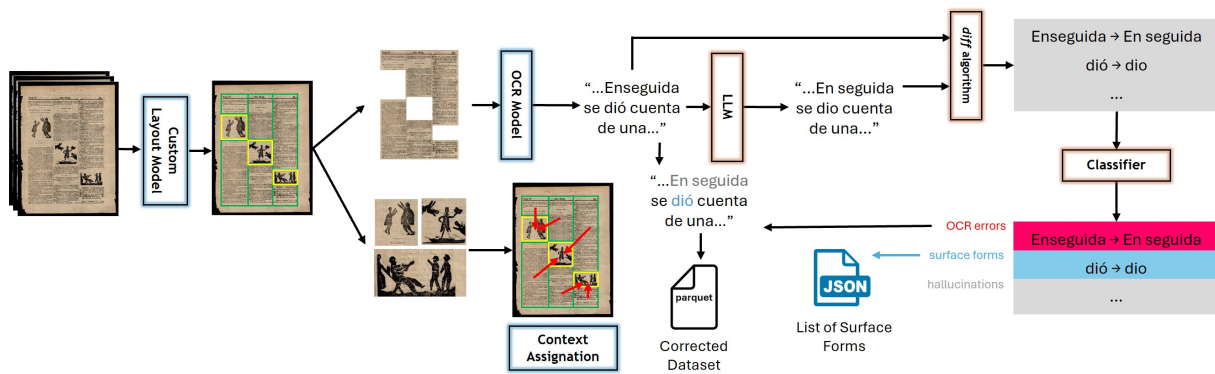


Figure 2: Overview of the full methodology pipeline. The blue components correspond to the Layout+OCR stage to get to digitized text, and the orange components correspond to the Post-OCR LLM Correction stage. The two outputs of the pipeline are the LatamXIX **Corrected Dataset** and the **List of Surface Forms**. The *Custom Layout Model* also extracts the images of the newspaper which are then assigned to the related texts (context). The final version of the text has the OCR errors corrected but not the surface forms, as they are part of the language.

format, such as JSON, as the *diff* algorithm produced shorter, more consistent, and less variable outputs. Additionally, this differentiation allows us to ignore any additions or deletions that result from LLM hallucinations, focusing instead on meaningful changes. An example of the original text, the corrected version, and the detected differences can be found in Appendix A, as well as the parameters chosen for this step.

3.3 Corrections Classification

Once the corrections are detected and isolated through the *diff* algorithm, the last step is to classify them. Still, first, it is important to state the main differences between the possible labels for each correction:

- **Surface form:** In linguistics, the term *surface form* (or word form) denotes the specific appearance of a word in a given context, contrasting with its lexical form, which pertains to its meaning (Sarveswaran et al., 2019). During the 19th century in Latin America, certain words were documented with variant spellings reflecting language shifts over time. It’s important to note that changes in surface forms do not necessarily alter the semantic content of the word, but rather represent orthographic modifications.
- **OCR error:** An OCR error, on the other hand, refers to every possible misread text from the real newspaper text. The OCR errors must be corrected but must be carefully separated newspaper linguistic "errors" that contribute to the linguistics of the time.

- **Hallucinations:** If none of the above is the case, the correction is an LLM hallucination or a translation to modern Spanish, which would be wrong, so these corrections must be omitted.

To enhance classification rule analysis, corrections were noted along with their frequency across the dataset to assess relevance. All corrections were converted to lowercase for effective grouping. Many corrections were reviewed and consolidated into a set of linguistic rules for categorization. This framework can be used to identify and analyze similar changes and classification rules in other languages and contexts. This paper presents a validated set of standardized rules and exceptions for classifying corrections in the LatamXIX dataset.

3.3.1 Accent changes

Corrections involving only accent changes (addition or removal) between the original and corrected texts refer mostly to **surface forms**, given the differences between 19th-century Spanish accent rules and modern ones (Montgomery, 1966). This includes varied accent expressions for the same word, such as "antes" sometimes written as "ántes". Surface forms pose problems for NLP tasks because, in Spanish, words without accents can have different meanings, such as "acepto" (present) and "aceptó" (past). Thus, for some NLP tasks, focusing on surface forms without accent changes may be preferable, which is another outcome presented in this paper.

Feature	Value
Size	~ 128MB
Rows	64,077
Words	~ 22M
Tokens	~ 28.7M
Newspapers	197
Years Range	1806 - 1899
Total Corrections	830,951
Surface Forms	37,492
Non-Accent Surface Forms	7,466
% of OCR Error Corrections	12.33%
% of Hallucinations Detected	77.96%

Table 1: Final Historical Ink: LatamXIX LLM Post-OCR corrected dataset

3.3.2 Specific changes

A set of letter-to-letter changes was extracted to represent key **surface words** and common **OCR errors**. For surface words, common changes include "y" for "i" or "g" for "j", e.g., "mui" for "muy" and "jeneral" for "general"; in fact, the connector "y" used to be written as "i" in most of the early 19th-century texts (Bouzouita and Gutiérrez, 2015). Common OCR errors include accent misreading or number confusion, such as "ó" read as "6" or "i" as "1". Appendix B shows a list of surface form changes.

3.3.3 Other letter-to-letter changes

When the number of letters in the original and corrected texts matches, changes generally refer to **OCR errors**, e.g., "la" misread as "In" or "señor" as "sefor".

3.3.4 Remaining changes

Corrections not fitting the preceding categories are challenging to classify as OCR errors or hallucinations, particularly with multiword corrections. A text similarity ratio was computed based on positional character matches between the original and corrected texts. This ratio, combined with the number of words in the corrected text and correction frequency, helped categorize corrections. For instance, "ascripeión" to "suscripción" had a ratio of 0.76, while "que" to "como" had a ratio of 0.0, effectively distinguishing most cases.

4 Results

Following the outlined steps, we produced the LatamXIX dataset, as shown in Table 1 and detailed in Appendix C, alongside a flexible LLM

OCR correction framework. This framework allows for easy interchange between datasets or LLMs, facilitating further research. We also compiled a list of 19th-century Latin American Spanish surface forms from newspapers and developed a general framework for detecting these forms in diverse contexts.

Old Spanish surface forms are particularly useful for semantic change detection, capturing meaning variations of specific words and aiding comparisons of their historical evolution across different periods and Spanish-speaking regions.

In terms of LLM post-OCR corrections, the system generated 830,951 corrections. However, a notable 78% of these were classified as hallucinations, indicating the model's tendency to generate incorrect or fabricated content when uncertain. Only 12% addressed actual OCR errors, reflecting the core objective of the framework. This gap highlights a key limitation of current LLM models in historical OCR correction, where distinguishing between genuine errors and hallucinations remains a challenge, especially in specialized datasets.

Moreover, due to Azure OpenAI's API content policy for the chosen LLM (GPT-4o-mini), 2,899 rows (4.52%) were excluded from processing because they contained content flagged as harmful, violent, or sexual. This limitation underscores the challenges content moderation policies pose when applying LLMs to historical texts. The percentage of flagged content provides insight into the prevalence of such material in 19th-century Spanish, offering valuable perspectives for comparative analysis with modern Spanish³.

5 Future Work

While the OCR correction using LLMs has progressed towards a more automated pipeline, a substantial portion of rule definition within the presented framework still requires manual professional input. To advance this process, future work should aim to enhance the automation of the rule-defining procedures. By reducing the reliance on human expertise, we can improve both the efficiency and accuracy of the OCR correction framework.

³The dataset, surface forms, and processing steps are available in <https://github.com/historicalink/LatamXIX>

6 Limitations

A significant limitation of this work is the reliance on manual evaluations for assessing OCR accuracy, as most evaluations and rule definitions were performed by experts. This manual process introduces subjectivity and limits scalability. The absence of a comprehensive automated evaluation method prevents more consistent accuracy assessments and restricts the ability to refine the framework based on objective metrics like Character Error Rate (CER).

7 Acknowledgements

We would like to thank the two anonymous reviewers from the EMNLP NLP4DH conference for their helpful feedback and suggestions.

References

- Alpha. 2023. [Newspaperbox Dataset](#).
- Chantal Amrhein and Simon Clematide. 2018. [Supervised OCR error detection and correction using statistical and neural machine translation methods](#). *Journal for Language Technology and Computational Linguistics*, 33(1):49–76.
- David M. Berry and Anders Fagerjord. 2017. *Digital Humanities: Knowledge and critique in a Digital age*. Polity Press.
- Miriam Bouzouita and Mara Fuertes Gutiérrez. 2015. [Spanish studies: Language and linguistics](#). *The Year's Work in Modern Language Studies*, 75:171–185.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. [Spanish pre-trained BERT model and evaluation data](#). In *PMLADC at ICLR 2020*.
- Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2017. [Icdar2017 competition on post-OCR text correction](#). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1423–1428.
- Ryan Cordell and David Smith. [The viral texts project](#). *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines (2022)*.
- James E. Dobson. 2019. *Critical Digital Humanities: The search for a methodology*. University of Illinois Press.
- Oceanic Exchanges. [The atlas](#). *Mapping the Histories and Metadata of Digitised Newspapers Collections Around the World. (2021)*.
- GrabadosXIX. 2023. [Grabados_Sample Dataset](#).
- National Endowment for the Humanities. [Chronicling america: Library of congress](#). *News about Chronicling America RSS*.
- Pierre-Carl Langlais. 2024. [Post-OCR-correction: 1 billion words dataset of automated OCR correction by llm](#). *Hugging Face*.
- Zoe LeBlanc. 2024. [More than keywords](#). *The American Historical Review*, 129(1):164–168.
- Daniel Lopresti. 2008. [Optical character recognition errors and their effects on natural language processing](#). In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, page 9–16. Association for Computing Machinery.
- Thomas Montgomery. 1966. [On the development of spanish y from "et"](#). *Romance Notes*, 8(1):137–142.
- Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. [Survey of post-OCR processing approaches](#). *ACM Comput. Surv.*, 54(6).
- Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. 2020. [Neural machine translation with BERT for post-OCR error detection and correction](#). In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, JCDL '20*, page 333–336, New York, NY, USA. Association for Computing Machinery.
- OCR. 2022. [OCR_project dataset](#).
- OpenAI. 2024. [GPT-4o mini: advancing cost-efficient intelligence](#).
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2019. [ICDAR 2019 competition on post-OCR text correction](#). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1588–1593.
- RSCOE. 2023. [Newspaper Dataset](#).
- Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2019. [Using meta-morph rules to develop morphological analysers: A case study concerning Tamil](#). In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 76–86, Dresden, Germany. Association for Computational Linguistics.
- SNSF and FNR. 2023. [Impresso - Media Monitoring of the Past II. Beyond Borders: Connecting Historical Newspapers and Radio](#).

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021a. [BART for post-correction of OCR newspaper text](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online. Association for Computational Linguistics.

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021b. [BART for post-correction of OCR newspaper text](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online. Association for Computational Linguistics.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. [Leveraging LLMs for post-OCR correction of historical newspapers](#). In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

M.E.B. Veninga. 2024. [LLMs for OCR post-correction](#).

A LLM Correction

A.1 Prompt

Below is the prompt used to request the LLM to correct the historical text extracted by the OCR model. This prompt remained unchanged for the correction of the entire dataset and was generated through manual trial and error, ensuring it was concise enough to accommodate the potential length of the text.

```
Dado el texto del siglo XIX
entre ``` , retorna únicamente
el texto corrigiendo los errores
ortográficos sin cambiar la
gramática. No corrija la
ortografía de nombres:
```

```
...
```

```
{text}
```

```
...
```

Equivalent to the following prompt in English:

```
Given the 19th-century text
between ``` , return only the text
with spelling errors corrected
without changing the grammar.
Do not correct the spelling of
names:
```

```
...
```

```
{text}
```

```
...
```

A.2 Example

The LLM response was successful for most of the texts except for some cases where Azure's Content Policy was triggered due to text content, and for very long texts where the model started to hallucinate the whole text. An example of an original text, its retrieved LLM correction, and all the changes detected by the diff algorithm is the following ([surface forms](#) and [OCR errors](#)) is:

- **Original:** La publicacion del Oso se harà dos veces cada se mana, y constará de un pliego en cuarto ; ofreciendo à mas sus redactores, dar los gravados oportunos, siempre que loexija el asuntode que trate. Redactado por un Num. 8. TEMA del Periodico. POLITICA MILITAR. OCTAVA SESION. Abierta la session á las dore y un minuto de la noche , 25 de Febrero de 1845 , con asistencia de todos los Señores Representantes, se leyó y aprobó la acta de la Asamblea anterior , ménos en lo tocante à la torre del Convento de Santo Domingo, punto que quedó para ventilarse en mejor ocasion. En seguida se dió cuenta de una nota del Ejecutivo , referente à que urjía la necesidad de organizar un Ejército ; pues decia el Excmo. Decano: - "Un poder sin bayonetas vale tanto como un cero puesto á la izquierda."
- **Corrected:** La [publicación](#) del Oso se [hará](#) dos veces cada [semana](#), y constará de un pliego en cuarto; ofreciendo [además](#) sus redactores, dar los [grabados](#) oportunos, siempre que [lo exija el asunto de](#) que trate. Redactado por un Num. 8. TEMA del [Periódico](#). [POLÍTICA MILITAR](#). OCTAVA [SESIÓN](#). Abierta la [sesión a](#) las [dos](#) y un minuto de la noche, 25 de Febrero de 1845, con asistencia de todos los Señores Representantes, se leyó y aprobó la acta de la Asamblea anterior, [menos](#) en lo tocante [a](#) la torre del Convento de Santo Domingo, punto que quedó para ventilarse en mejor [ocasión](#). [Enseguida](#) se [dio](#) cuenta de una nota del Ejecutivo, referente [a](#) que [urgía](#) la necesidad de organizar un Ejército; pues [decía](#) el Excmo. Decano: - "Un poder sin bayonetas vale tanto como un cero puesto [a](#) la izquierda."

B Specific Surface Form Changes

For the surface form extraction from the texts and its differentiation from OCR errors and LLM hallu-

cinations, a set of surface form changes was constructed for 19th-century Latin American Spanish. The complete set of known changes with an example for each case is presented in Table B1.

Change	Example
á ↔ a	hara → hará
é ↔ e	fué → fue
í ↔ i	decia → decía
ó ↔ o	ocasion → ocasión
ú ↔ u	ningun → ningún
i ↔ y	mui → muy
j ↔ g	jente → gente
v ↔ b	gravado → grabado
s ↔ x	espiró → expiró
j ↔ x	méjico → méxico
c ↔ s	faces → fases
s ↔ z	dies → diez
z → c	doze → doce
q → c	quatro → cuatro
n → ñ	senor → señor
ni → ñ	senior → señor
k → qu	nikel → níquel
k → c	kiosko → quiosco
ou → u	boulevard → bulevar
s → bs	suscriciones → suscripciones
c → pc	suscriciones → suscripciones
s → ns	trasportar → transportar
t → pt	setiembre → septiembre
rt → r	libertar → liberar
r ↔ rr	vireinato → virreinato
...lo → lo ...	cambiólo → lo cambió
...se → se ...	acercóse → se acercó

Table B1: Set of Surface Form change rules to extract them from the LatamXIX dataset

C Dataset Overview

A more specific overview of the dataset is described in Figure C1 and Table C1.

Country	Presence (%)
Mexico	49.59%
Argentina	21.23%
Colombia	12.53%
Peru	8.43%
Chile	6.39%
Panama	0.83%
Venezuela	0.52%
Uruguay	0.17%
France	0.16%
Ecuador	0.09%
Spain	0.06%

Table C1: LatamXIX dataset presence distribution grouped by country

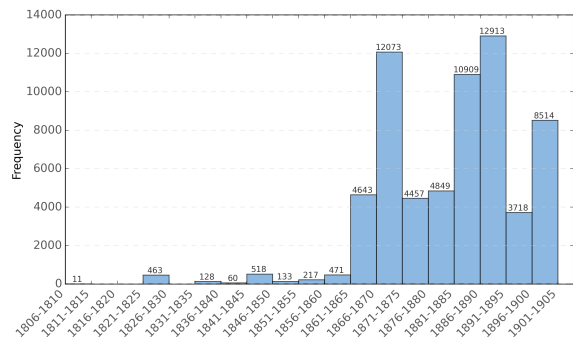


Figure C1: LatamXIX dataset decade distribution

Canonical Status and Literary Influence: A Comparative Study of Danish Novels from the Modern Breakthrough (1870–1900)

Pascale Feldkamp, Alie Lassche, Jan Kostkan, Márton Kardos,
Kenneth Enevoldsen, Katrine Baunvig and Kristoffer Nielbo

Aarhus University

{pascale.moreira, kenneth.enevoldsen}@cc.au.dk,
{a.w.lassche, jan.kostkan, martonkardos, baunvig, kln}@cas.au.dk

Abstract

We examine the relationship between the canonization of Danish novels and their textual innovation and influence, taking the Danish Modern Breakthrough era (1870–1900) as a case study. We evaluate whether canonical novels introduced a significant textual novelty in their time, and explore their influence on the overall literary trend of the period. By analyzing the positions of canonical versus non-canonical novels in semantic space, we seek to better understand the link between a novel’s canonical status and its literary impact. Additionally, we examine the overall diversification of Modern Breakthrough novels during this significant period of rising literary readership. We find that canonical novels stand out from both the historical novel genre and non-canonical novels of the period. Our findings on diversification within and across groups indicate that the novels now regarded as canonical served as literary trendsetters of their time. To ensure reproducibility, code and raw data are available at <https://github.com/centre-for-humanities-computing/memo-canonical-novels>.

1 Introduction

At the beginning of the 21st century, the Danish government published an Educational Canon (*Undervisningskanon*) and a Cultural Canon (*Kulturkanon*) in an attempt to improve knowledge of Danish literature among the population, and to standardize school syllabi (Harbild et al., 2004). Both canons were met with criticism, and the canon debate flared up in full force – a development similar to, e.g., the Netherlands and Germany (Mai, 2016; Smid, 2022). Much of the criticism in Denmark was of the canon being unrepresentative and lacking diversity, including but one female author at the time (Fibiger, 2004).

The Danish canon debate echoes a central issue in literary scholarship, dating at least to the ‘canon

wars’ of the 1980s (Guillory, 1995; Witt, 2000): the critique that the canon is a top-down, contemporary construct that does not accurately reflect works’ historical significance, their impact on readers, or the breadth of literary production of a period. Still, advocates for the ‘canon’ being a meaningful term¹ argue that canonical works exhibit an enduring literary value and distinguish themselves by their lasting influence and innovation at the textual level (Bloom, 1995; Van Peer, 2008).

In this paper, we examine two hypotheses reflecting these polar stances on the canon: that canonical works are top-down and present day constructs, so that they would not stand out textually from their contemporaries (H1); and that canonical works distinguish themselves by textual novelty and literary influence, in which case we would expect them to show an impact on their literary field (H2).

To gauge whether books that are (today) considered canonical exhibit these distinguishing traits – textual innovation and literary impact and whether this resonates in what was published after – we compare canonical novels to novels that did not make it into the different constructions of ‘canon’ in a case study of Danish novels written in the Modern Breakthrough era. This late 19th century period is particularly suited as a case study of canon/non-canon dynamics for three reasons: 1) It allows us to examine the status of canonical works within their historical context. 2) It allows us to work with a complete corpus of the literary production of a time period – albeit limited. This is particularly significant because attempts to address issues of canon representativity often face the challenge of not catching the ‘dark numbers’ of literary production – the extensive numbers of titles forgotten or overlooked (i.e., the ‘great unread’ (Moretti, 2000)).

¹Many literary scholars argue the opposite, that the canon should be rethought or revised (von Hallberg, 1983), and that terms like ‘classics’ belong to the “precritical era of criticism itself” (Guillory, 1995).

Lastly, 3) the Modern Breakthrough era is a period of significant political and cultural upheaval, where we would expect to see literary innovation.

The Modern Breakthrough (1870-1900) – in Danish, *det Moderne Gennembrud* – marked a significant shift towards realism and naturalism, diverging from the romantic and idealistic styles that characterized the preceding period (D’Amico, 2016). Spearheaded by the influential critic Georg Brandes,² this era brought as much cultural as social change (Bjerring-Hansen and Wilkens, 2023). Literature of the period was unprecedented in emphasizing social issues, individualism, and a scientific approach, advocating for art to reflect and critique society (Mai, 2022). Moreover, the period saw a great rise in the number of literary publications (Bjerring-Hansen and Jelsbak, 2010), as well as an incline in previously underrepresented voices: journalists, teachers, and female authors published more (Bjerring-Hansen and Wilkens, 2023).

The current study focuses exclusively on Modern Breakthrough novels, which is not just a methodological choice to increase comparability,³ but also recognises the novel’s central position in the literary field of the late 19th–20th century. In this period, the novel expanded its reach to a broad and diverse readership, whereas poetry largely catered to a limited elite audience (Levine, 2008; Bjerring-Hansen and Jelsbak, 2010). This democratization of literature, coupled with the novel’s generic capacity to reflect the complexities of a rapidly changing society⁴ – social, political, and personal – renders it the period’s most dynamic and malleable genre. It is a genre in which we expect much – even short-term – development in this period. As both a popular and prestigious genre, the novel also reflects the evolving tensions between canonical authority and popular appeal, making it an unparalleled document for tracing literary influence and its directionality in the period.

This paper is structured as following: Section 2

²Brandes’ first Copenhagen lecture of the series “Hovedstrømninger” (1871), and the publication of J.P. Jacobsen’s *Mogens* (1872) are often pinpointed as the start of the Modern Breakthrough (Bjerring-Hansen and Rasmussen, 2023).

³This choice is naturally also restricted by data availability, as a corpus spanning the whole population (i.e., covering the ‘dark numbers’ of literary production is a rare resource).

⁴The ability to reflect social reality is an often highlighted generic trait, as in the seminal *The Rise of the Novel* of Watt (2001), describing the novel as ‘truth to individual experience’. Similarly, Armstrong (1987) suggested that the 19th century novel reflected societal upheaval but was also an important instrument of change in bringing the middle class to light.

contains a discussion of related work on canonicity, literary innovation and influence, and the Danish literary context. In Section 3, we describe the dataset and annotations. Our methodological pipeline is described in Section 4, and includes the creation of document embeddings using both a multilingual model and TF/IDF, clustering methods for validating embeddings, and measuring diachronic change to explore how the canon and non-canon evolve over time. The results are presented in Section 5, followed by a discussion in Section 6. We finish with concluding remarks in Section 7, and a discussion of the limitations of this study (Section 8).

2 Related Work

2.1 Textual profile of canonical works

The discussion about canon has been torn between two extreme poles, where canonicity is either seen as something conferred ‘from above’ or as signaling the excellence of particular works ‘from below’ in terms of text-intrinsic features (Bloom, 1995). Recent studies show a nuanced take on the debate: while they show that text-extrinsic features⁵ might be good predictors of canonicity (Brottrager et al., 2021), canonical works also appear to have a unique textual profile compared to non-canonical works (Barré et al., 2023; Brottrager et al., 2021; Porter, 2018). Beyond the binary distinction (canon/non-canon), canonical works exhibit textual profiles different from other types of excellence categories in literature, e.g., bestselling or prize-winning novels (Bizzoni et al., 2024; Wu et al., 2024). They have been found to have a denser nominal style (Wu, 2023) and lower readability, elicit higher LLM perplexity, and show more unpredictable sentimental dynamics (Bizzoni et al., 2024).

The axis along which canonical works are analyzed could be termed ‘stylistic difficulty’. Here, traditional linguistic metrics and information theory have been employed to show that texts with greater literary prestige tend to exhibit higher levels of reading difficulty (Algee-Hewitt et al., 2016; Bizzoni et al., 2024; Wu et al., 2024), than more ‘popular’ works of literature which use a more accessible language, and find a broader audience (Bizzoni et al., 2023). However, few studies go beyond features of linguistic and stylistic complexity in

⁵I.e., cultural, political or market traits, as in Wang et al. (2019).

examining the canon, although some have shown that both sentiment and semantic profiles may be good predictors of more popular literature (Maharjan et al., 2018; Bizzoni et al., 2024).

Since literature is clearly a multidimensional phenomenon, we ideally take all these textual levels into account when we try to grasp the difference between canon and non-canon. Therefore, we make use of document embeddings, which are able to capture text characteristics at various levels, including stylistics and semantics (Wang et al., 2023; Terreau et al., 2024; Reimers and Gurevych, 2019).

Moreover, while textual metrics are generally used to predict a modern label (e.g., what has been shown to sell well/has become canon), few studies have looked into the dynamics of the literary field within the period itself (Brottrager et al., 2022). Although authors like Henrik Pontoppidan are regarded as canonical and influential today, it is uncertain whether their exceptional status was equally recognized by their contemporaries. Thus, the issue of canonicity is closely tied to the concepts of intertextuality and literary influence, which have traditionally focused on how individual authors were shaped by their predecessors (Bloom, 1975, 2011; Bassnett, 2007). If canonicity can be viewed as a marker of reception, we must consider how latent this reception was. In the context of this study, we are interested in the direction of literary influence, specifically whether books that become canonical influence subsequent novels or if they adapt to the overall novel production.

2.2 Canon and popular literature in the Modern Breakthrough

Reading audiences grew significantly during the Modern Breakthrough, and a more differentiated selection of literature became available to more and better readers than before (see also the increase of novels published in Appendix A) (Bjerring-Hansen and Jelsbak, 2010; Hertel Hans, 1983).⁶ In the period, we also see an intellectual disdain for the ‘popular novel’, what Brandes spoke of as ‘døgnlitteratur’ (ephemeral literature, or literature ‘of the day’) (Brandes, 1877). *Døgnlitteratur* included, for example, the historical novel such as by Walter Scott, who had an enormous influence on the Dan-

⁶Beyond the growth of novel readership in the period, which Bjerring-Hansen and Wilkens (2023) call a ‘reading explosion’, Danish daily press also saw a great increase in this period, from 36 newspapers in 1847 to 156 in 1914 (Bjerring-Hansen and Wilkens, 2023).

ish and European literary field in the time preceding the Modern Breakthrough (Munch-Petersen Erland, 1978; Lukács, 1964).⁷ As a match for the popularity of translated Scott novels (Munch-Petersen Erland, 1978), in the Danish context, especially B.S. Ingemann should be foregrounded. Ingemann had a diverse audience – from sailors to the (Sorø) academy – and received the same disdain from the intellectual elite in the period of the Modern Breakthrough as Scott (Bjerring-Hansen and Rasmussen, 2023). While the scorn of the popular novel was itself not a new phenomenon – also present in the reception of Ingemann (Martinsen, 2012) – it was in the Modern Breakthrough accompanied by a decline in the historical novel genre (Bjerring-Hansen and Rasmussen, 2023), and a rise in what Bjerring-Hansen and Wilkens (2023) have broadly called the ‘realist novel’, pitching the two types of novels starkly against each other (Bjerring-Hansen and Rasmussen, 2023).

However, this polarization within the genre and the dynamics of trends and innovation in the novel of this period are less explored – a period where the appearance of the Modernists in Danish literature coincided with the decline of the previously very viable popular genre of the historical novel at a time at which the demand for popular literature was on the rise.

3 Data

Our dataset consists of a collection of 838 original Danish and Norwegian novels (1870-1900), with connected metadata, e.g., number of pages, book prizes, and publishing house.⁸ Previously, Bjerring-Hansen and Rasmussen (2023) tagged the corpus for whether a work is a historical novel or not. The corpus consists of all original first-edition novels published by Danish publishers in the period.⁹

As we sought to examine the relationship be-

⁷Moretti has also shown how the historical novel à la Scott gained a predominant position in the literary field 1740-1840, marginalizing older genres (Moretti, 2007).

⁸All novels, including the ones written by Norwegian authors, were published in the Danish language and at Danish publishing houses.

⁹The MiMe-MeMo corpus was compiled by Jens Bjerring-Hansen, Philip Diderichsen, Dorte Haltrup, and Nanna Emilie Dam Jørgensen, based on the Danish book index (*Dansk Bogfortegnelse*). It indexes all publications (1830-), including novels by Norwegian authors at Danish publishers. Creators excluded everything not novels (e.g., short story collections). For details, see Bjerring-Hansen et al. (2022). Version 1.1 (used in the present study) is accessible at: <https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1>.

tween today’s canonized novels from the Modern Breakthrough and the overall production of the period, we added a tag that informs us about the canonicity of the work. To compare the canons defined by a government-designated committee – which do not include Norwegian authors – with a canon that we assume to be created from a literary expert point of view (and less driven by a political agenda), we create a second canon that includes novels that are listed and mentioned in the lemma ‘det moderne gennembruds litteratur’ of the encyclopedia *Den Store Danske*.¹⁰ We thus added the following tags to the novels in our corpus:

- **CE Canon:** Cultural/Educational Canon, referring to novels whose titles are included in the Cultural Canon, or whose author is included in the Educational Canon.
- **LEX Canon:** Lexicon Canon, referring to novels that were not included in the Educational Canon and Cultural Canon, but whose author is mentioned in the lists of novels and novellas in the ‘det moderne gennembruds litteratur’ lemma of *Den Store Danske*.
- **E Canon:** Extended Canon, referring to all novels that are included in CE Canon and/or LEX Canon.¹¹
- **Other:** Other, referring to the novels that are neither tagged as historical, nor included in one of the canons.

Statistics of the corpus and every category can be found in Table 1.¹²

4 Methods

We developed a methodological pipeline consisting of the following steps:

1. Choosing embedding model. We test four embedding models to decide on the one best suited for our task and corpus. We test these using a weighted average between a historic clustering

¹⁰See https://denstoredanske.lex.dk/det_moderne_gennembruds_litteratur. We are aware that the Educational and Cultural Canon and the *Den Store Danske* lemma include more genres than the novel. This paper focuses specifically on the canonical reputation of the novel.

¹¹Note that some tags overlap, so that we tag as historical–canon in the following visualizations anything that was both tagged historical and was in either of the canons.

¹²An extended dataset (with added tags) is available on HuggingFace: <https://huggingface.co/datasets/chcaa/memo-canonical-novels>.

	titles	authors
Corpus	838	371
Cultural/Educational Canon	36	6
Lexicon Canon	110	19
Extended Canon	114	21
Historical Novels	65 (8)	19 (4)
Other	667	335

Table 1: Statistics on the corpus. Note that there is overlap between the categories: there are titles that are both in the Cultural/Educational Canon and the Lexicon Canon. The numbers between brackets in the Historical Novels category refer to titles that are tagged as a historical novel, but also included in one of the canons.

	\bar{x} SEB	Historical	SoI
Number of Datasets →	24	1	25
Models ↓			
m-e5-large-instruct	66.65	40.10	53.38
m-e5-large	<u>60.69</u>	27.66	44.18
DFM-large	55.14	35.13	<u>45.14</u>
MeMo-BERT	36.85	<u>35.38</u>	36.12

Table 2: The performance of encoder models on the Scandinavian Embedding Benchmark (SEB) tasks and on the custom historical task. The Score of Interest (SoI) reflects the model’s average score across tasks. The highest score is in bold, and the second highest is underlined.

task and the Scandinavian Embedding Benchmark (SEB)¹³ (Enevoldsen et al., 2024) to get a model performing well generally and across historical documents. The performance of the four models can be found in Table 2. For our models, we use the MeMo-BERT trained on Danish and Norwegian historical documents (Al-Laith et al., 2024), the best-performing Danish sentence encoder DFM-large (Enevoldsen et al., 2023) along with the two best-performing open-weight¹⁴ models on SEB, m-e5-large as well as its prompt-based version m-e5-large-instruct (Wang et al., 2024b). Prompt-based models allow for adaptation of the embedding space depending on the use case and have been shown to improve performance significantly (Muennighoff et al., 2023; Enevoldsen et al., 2024; Wang et al., 2024b) as seen in Table 2 we also find this to be the case. For the prompt-based model, we used the instruction “Identify the author of a given passage from historical Danish fiction” for evaluation of the historical task. For readability,

¹³We use the latest version of SEB (v0.13.6).

¹⁴We avoid using commercial APIs to ensure reproducibility.

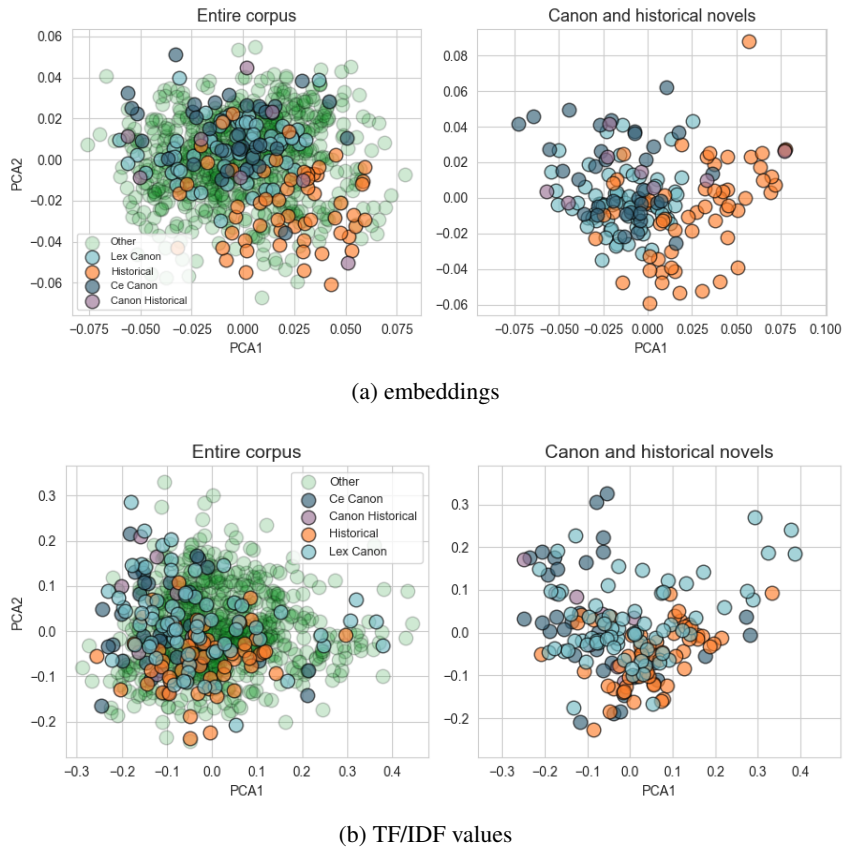


Figure 1: PCAs of the entire corpus (*left*) and the canonical and historical novels (*right*), based on embeddings and TF/IDF values. Note that canon and historical groups cluster more in the PCA based on embeddings.

we short the model names. For full model references along with revision, see [Appendix F](#). We present the construction of the historic task in [Appendix G](#). We continued with the best performing model, m-e5-large-instruct.

2. Creating document representations. We create two types of document representations: our main approach is creating *semantic embeddings*, while we use *lexical embeddings* to validate our semantic embeddings.

- **Semantic embeddings.** We slice each novel into chunks of the same size.¹⁵ Afterwards, we create embeddings for every chunk with the m-e5-large-instruct model, using the same prompt as in the previous step. The average of all document embeddings of one novel is used as a representative embedding for that novel.
- **Lexical embeddings.** After pre-processing the documents (lowercasing, removing punctuation), we create a TF/IDF representation

¹⁵Since the maximum chunk size includes the length of the prompt, we use a chunk size of $512 - 87 = 425$ tokens.

of each novel using sklearn ([Pedregosa et al., 2011](#)).

3. Clustering embeddings for method validation.

We validate our method by clustering the obtained document embeddings using different measures and visualizations, including dendrograms (see [Appendix B](#)) and a PCA as implemented in sklearn ([Pedregosa et al., 2011](#)). We use PCA as it preserves the global structure of the embedding space.

4. Measuring diachronic change. We use intra- and inter-group (cosine) similarity to measure how the canon and non-canon evolve over time and how they influence each other.

5 Results

5.1 Validation of embeddings

We perform clustering methods on the two different types of embeddings to verify the novel distribution based on semantic and lexical features. The two PCAs in [Fig. 1a](#) are based on the semantic embeddings of the novels. The left PCA shows that overall, the novels that are tagged as canonical and/or historical (colored blue, purple, and orange)

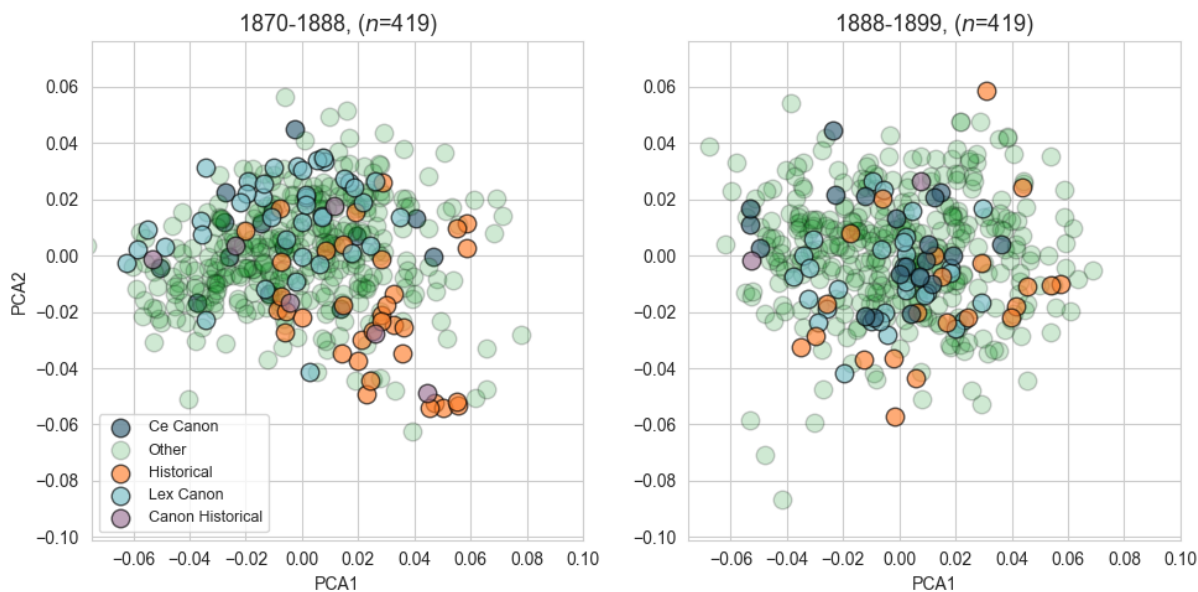


Figure 2: PCA’s of (*left*) the first half of the corpus (1870-1888) and (*right*) the second half of the corpus (1888-1899), with novels ordered chronologically.

largely overlap with the rest of the corpus (colored green). This tells us that, at first sight, their semantic style does not differ significantly from the overall literary production during the Modern Breakthrough. When we only look at the canonical and historical novels, as visualised in the right PCA, we see a distinction between the historical novels (orange) and the canonical novels. This can also be confirmed in the dendrogram in [Appendix B](#) (Fig. 6a), in which the clear orange cluster of historical novels at the right side of the plot suggests that they share similarity in semantic space.

When we compare these figures with the ones based on the TF/IDF values, we see some interesting differences. The PCAs (Fig. 1b) do not show as clear clusters of canonical novels and historical novels, and the same goes for the dendrogram (Fig. 6b in [Appendix B](#)). The works of the earlier mentioned author Pontoppidan for example do not cluster on lexical style, and only parts of the historical novels cluster together, while the rest is spread out over the other branches. It suggests that our semantic embeddings go beyond lexical features. This is in line with previous results ([Enevoldsen et al., 2024](#)) indicating that document embeddings primarily capture semantics and, e.g., cannot differentiate between correct/incorrect word order.

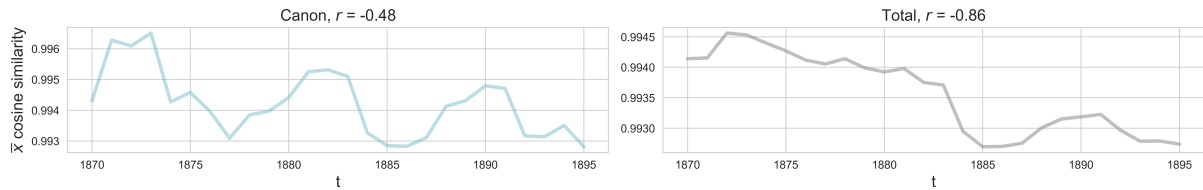
5.2 Diachronic change

A diachronic comparison between semantic embeddings of the first 419 novels in our corpus (pub-

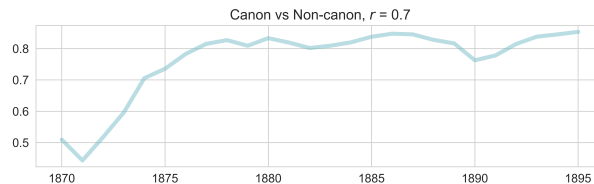
lished between 1870 and 1888) and the latter 419 novels (published between 1888 and 1899) shows a difference over time in the relationship between canonical novels and the rest of the corpus (see Fig. 2). While the left PCA shows that canonical novels (colored light and dark blue) cluster on the upper-left side of the green cluster (with non-canonized novels) in the period 1870-1888, the canonized and non-canonized novels show much more overlap in the later period. It suggests that the early Modern Breakthrough works that became canonical differed semantically from the overall production of that time, while the later canonical works were more similar to the other novels of that time.

To verify this potential diachronic change, we compute the mean embedding of canonical novels and non-canon novels for each rolling window (window size of 4 years) across the 30-year period and plot the cosine similarity between the two groups for each window (Fig. 3b).¹⁶ We see a ro-

¹⁶Due to the discrepancy in group sizes (e.g., canon vs. non-canon) and the overall skewed distribution of our corpus (see Fig. 5 in [Appendix A](#)), we used simulation methods to compare cosine similarity across time windows. For each window, we simulated 1,000 Gaussian distributions of cosine similarity for each group based on their respective means and standard deviations. The overall mean of these 1,000 runs was used for each group in the comparison. For intergroup comparisons, we employed the same approach by simulating the mean embedding of each group (1,000 runs per window) and then calculating the average cosine similarity between the groups’ embeddings across runs for each window. To ensure



(a) Intra-group similarity across time, using the mean cosine similarity of books in a rolling window ($s = 4$, $step = 1$) over the years for each category – all books and our extended canon definition. Spearman’s ρ of the correlation between year and mean intragroup similarity (on top), $p < 0.01$. For each rolling window, group size > 2 . Note that both groups tend to get more internally diverse across time, though canon books less so and less consistently.



(b) Inter-group similarity across time: The cosine similarity between the average canon and non-canon embedding per rolling window ($s = 4$, $step = 1$). Note that a correlation with time persists after 1875 ($\rho .42$).

Figure 3: Intra-group similarity and inter-group similarity across time.

bust positive correlation with time (Spearman’s $\rho .7$, $p < 0.01$),¹⁷ suggesting that canonical novels become increasingly similar to non-canonical novels. This positive correlation is partly due to the steep increase in inter-group similarity before 1875. Nevertheless, the correlation with time persists after that year ($\rho .42$). The flattening of the curve might suggest that certain subgenres continue to disappear (like the decline of the historical novel), or it might reflect how the entire literary field is more standardized from 1875 and onwards.

When we look at the intra-group similarity of all canonical works – using the same rolling window to extract intra-group cosine similarity over time – we see a decrease in similarity internally in the canon group, suggesting that the canon group becomes more internally diverse over time. The same trend, though slightly stronger, can be observed in the corpus as a whole (Fig. 3a).¹⁸ Note that these correlations of intra- and inter-group similarity over time hold regardless of which model (among those tested) is used to create embeddings (see Table 3 in Appendix C).

To detect whether the canon moves towards the non-canon over time, or the other way around, we

results are not skewed because we assume a normal distribution, we also tested a bootstrap sampling, which yielded similar results.

¹⁷Due to our simulation approach, results may vary slightly for each run, so correlation coefficients with time should be taken as estimations rather than precise values.

¹⁸For further validation of the limited range of cosine similarity values in our study, see Figure 7b in Appendix D.

gauge the directionality of both groups. We split the corpus in two equal parts, in the same way as done for Fig. 2. The result consist of four subsets: early non-canonical novels, early canonical novels (both pre-1888), late non-canonical novels and late canonical novels (both post-1888). We have plotted all novels in one PCA (Fig. 4), using colors to distinguish between the four subsets. We fit the mean embedding of every subset to the same PCA. The resulting plot shows that the late non-canon has moved up in the direction of the early canon, suggesting that the novels that today have a canonical status, behaved as trendsetting novels at the time. In Appendix E, we include an alternative version of this PCA, using a rolling window size of 5 years (step 1) to show that the non-canon moves towards the canon.

6 Discussion

Seeking to validate what our embeddings capture, we compared canonical and historical novels both in terms of embedding space and lexical similarity (TF/IDF). The fact that these two groups of novels – canonical and historical – cluster differently at the level of embeddings is interesting for two reasons: firstly, it suggests a maintained coherence of the historical genre in this period, although it was in decline (Bjerring-Hansen and Rasmussen, 2023). Secondly, since this is not an equal comparison – one group being a genre and the other a category spanning a diversity of novels – we find that

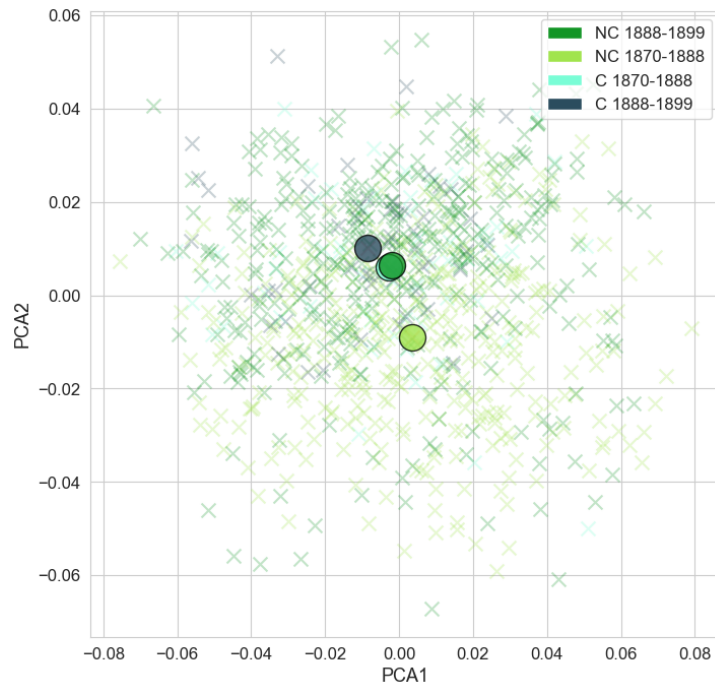


Figure 4: Positions of the mean embedding of the canon/non-canon groups over two time-periods (1870-1888, 1888-1899). Note that the later period non-canon mean seems to have moved closer to mean canon position. The PCA was fitted to all embeddings, then mean embeddings per group per period were fitted into the same PCA.

the clustering must suggest that our embeddings capture more than thematic elements, which more usually distinguish more popular genres (Moreira and Bizzoni, 2023).

Moreover, our clustering experiments show that different constructions of the canon, i.e., the more political, smaller canon, and the expert canon which includes non-Danish authors to a higher degree, do not seem to differ significantly: both versions include novels that stand out based on their semantic features. The textual features of the small selection of novels included in the Cultural Canon and the Educational Canon are not different from those of the novels that are included in what we call the Lexical Canon. This can be derived from the way in which novels from both groups have close proximity to each other and are mixed in the dendrograms (Figs. 6a and 6b).

For our main result, we show that the Modern Breakthrough novels that have a canonical status today show traits at the textual level which suggest their innovation and distinction from their non-canonical contemporaries, contra to our H1. This ties in with what is shown in earlier studies on the literary canon: canonical novels have text-intrinsic features that distinguish them from other novels (Brottrager et al., 2022; Bizzoni et al., 2024; Wu,

2023). Moreover, our results suggest that canonical novels behaved innovatively, introducing characteristics in the Modern Breakthrough era, which resonated in the literary production that came after – non-canonical novels in a sense tracing after the canonical novels. In a diachronic comparison, we see that non-canonical novels adapt to the canon, likely copying the innovative themes and style that these trendsetters introduced. This supports our H2. It makes the current study – to the best of our knowledge – the first that uses embeddings to show the relationship between canonical and non-canonical novels in terms of innovation and influence.

Furthermore, the decreasing intra-group similarity of both the canon and non-canon reflects the diversification of the literary field during the period, as outlined by (Bjerring-Hansen and Jelsbak, 2010; Hertel Hans, 1983). A larger variation of novels saw the light of day, and a more diverse selection of novels became canonical, reflected in canonical novels’ increasing internal diversity. While this study confirms the textual innovation of canonical novels, it is possible that there are novels that show the same textual profile as the canon but that did not get canonical status. Future research should provide more insight into these potentially innovative

but today lesser known novels. Moreover, while we argue that the canonical status of novels might be related to their textual profile, we acknowledge that text-extrinsic features could also have played a role here. Future work could, therefore, explore the relationship between the canonical status of a novel and features such as the price of the book, and the publishing house – something the current dataset allows for.

A last note concerns the directionality of the canon and non-canon works, where our results suggest that works in the canon group may have acted as literary trendsetters. Further research, employing more sophisticated methods to gauge causality is needed to confirm our suggestion. Moreover, future work could compare the embedding spaces of canon and non-canon groups to the embedding space of non-fiction texts, as the latter may serve as a useful reference point for assessing the movement and direction of fiction works.

7 Conclusion

We have examined the relationship between the canonization of Danish novels and their textual innovation and influence, taking the Danish Modern Breakthrough era (1870–1900) as a case study. We created embeddings of the 838 novels in our corpus, and used a custom historic clustering task to decide on the best suited model for our task and corpus, which turned out to be the multilingual m-e5-large-instruct model. We validated our embeddings by creating a TF/IDF representation of each novel. Our results show that the embeddings capture semantic features and go beyond lexical features: historical novels and canonical novels cluster differently. Inter-group similarity shows that the similarity of canonical and non-canonical novels increases over time, while at the same time, intra-group similarity decreases, indicating that the canon group as well as the overall novel production becomes more internally diverse over time. We finally show that the non-canon moves towards the canon, suggesting that non-canonical novels adapt to the canon, possibly copying the innovative themes and style of these trendsetters.

8 Limitations

Prompts for embeddings: This work utilizes the prompt-based embedding model m-e5-large-instruct, and thus, it is likely that notably different results could have been obtained by changing the

prompt. We examine this further in [Appendix H](#).

Occurrence within training data: Canonical works are more likely to appear online or outside their original context due to their popularity. This could lead to differences in embeddings when using models trained on large web-based data sources simply because paragraphs from these novels may appear in varied contexts within the training data. However, we consider this influence to be minor, as historical novels likely represent only a small fraction of online discourse. This is especially the case for the multilingual embedding model used, where Danish likely comprises only a tiny fraction of the training data. Ideally, the training data should be examined to ensure this influence is not significant. However, this approach is often unfeasible, as pre-training data for these models is typically unavailable, and exploring it would require extensive computational resources. Additionally, the fact that historical canon has often been rewritten further complicates such efforts.

Canon definition: the concept of canonicity is inherently vague and subject to various interpretations. Our canon definition and our binary classification of canonical works may oversimplify a concept that may be better represented as a continuous variable ([Brottrager et al., 2022](#)). Our rationale in using two ideal classes (canon/non-canon) was to get an estimate of the difference between them, though it should be noted that the transition between them may be more fluid than it is represented here.

Acknowledgements

The authors of this paper were supported by grants from the Carlsberg Foundation (grant title: *The Golden Array of Danish Cultural Heritage*) and the Aarhus Universitets Forskningsfond (grant title: *Golden Imprints of Danish Cultural Heritage*). Part of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Herscovich. 2024. [Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources*

- and Evaluation (LREC-COLING 2024), pages 4811–4819, Torino, Italia. ELRA and ICCL.
- Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. *Canon/Archive. Large-scale Dynamics in the Literary Field*. Stanford Literary Lab.
- Nancy Armstrong. 1987. *Desire and domestic fiction: a political history of the novel*, nachdr. edition. Oxford Univ. Press, New York.
- Jean Barré, Jean-Baptiste Camps, and Thierry Poibeau. 2023. [Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature](#). *Journal of Cultural Analytics*, 8(3).
- Susan Bassnett. 2007. [Influence and Intertextuality: A Reappraisal](#). *Forum for Modern Language Studies*, 43(2):134–146.
- Yuri Bizzoni, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2024. [Good Books are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality](#).
- Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023. [Good reads and easy novels: Readability and literary quality in a corpus of US-published fiction](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 42–51, Tórshavn, Faroe Islands. University of Tartu Library.
- Jens Bjerring-Hansen and Torben Jelsbak. 2010. *Boghistorie*, 1 edition. University Press, Aarhus.
- Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. [Mending Fractured Texts. A heuristic procedure for correcting OCR data: 6th Digital Humanities in the Nordic and Baltic Countries Conference, DHNB 2022](#). In *CEUR Workshop Proceedings*, volume 3232, pages 177–186, Uppsala, Sweden.
- Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. [Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne genembrud som case](#). *Passage - Tidsskrift for litteratur og kritik*, 38(89):171–189.
- Jens Bjerring-Hansen and Sebastian Ørtoft Rasmussen. 2023. [Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne genembrud som case](#). *Passage - Tidsskrift for litteratur og kritik*, 38(89):171–189. Number: 89.
- Jens Bjerring-Hansen and Matthew Wilkens. 2023. [Deep distant reading: The rise of realism in Scandinavian literature as a case study](#). *Orbis Litterarum*, 78(5):335–352. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/oli.12396](https://onlinelibrary.wiley.com/doi/pdf/10.1111/oli.12396).
- Harold Bloom. 1975. *The Anxiety of Influence: A Theory of Poetry*. Number 426 in A Galaxy Book ; GB. Oxford Univ. Press, London.
- Harold Bloom. 1995. *The Western Canon: The Books and School of the Ages*, 1st riverhead ed edition. Riverhead Books, New York, NY.
- Harold Bloom. 2011. *The Anatomy of Influence: Literature as a Way of Life*, 1 edition. Yale University Press, New Haven.
- Georg Brandes. 1877. *Hovedstrømninger i det 19de Aarhundredes Litteratur : Forelæsninger holdte ved Københavns Universitet 1871-1887. : Emigrantlitteraturen*. 2. udgave. København, Gyldendal.
- Judith Brottrager, Annina Stahl, and Arda Arslan. 2021. [Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features](#). In *CEUR Workshop Proceedings*, pages 195–205, Antwerp, Belgium. CEUR.
- Judith Brottrager, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin. 2022. [Modeling and predicting literary reception](#). *Journal of Computational Literary Studies*, 1(1):1–27.
- Giuliano D’Amico. 2016. [Modern Breakthrough](#). In *Routledge Encyclopedia of Modernism*, 1 edition. Routledge, London.
- Kenneth Enevoldsen, Lasse Hansen, Dan S. Nielsen, Rasmus A. F. Egebæk, Søren V. Holm, Martin C. Nielsen, Martin Bernstorff, Rasmus Larsen, Peter B. Jørgensen, Malte Højmark-Bertelsen, Peter B. Vahlstrup, Per Møldrup-Dalum, and Kristoffer Nielbo. 2023. [Danish foundation models](#).
- Kenneth Enevoldsen, Kardos Marton, Niklas Muenighoff, and Kristoffer L. Nielbo. 2024. [The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding](#).
- Johannes Fibiger. 2004. [Kampen om litteraturhistorien. Om veje ind i og ud af litteraturen](#). *Bogens Verden. Tidsskrift for kultur og litteratur*, (6).
- John Guillory. 1995. *Cultural Capital: The Problem of Literary Canon Formation*. University of Chicago Press.
- Steen Harbild, Stefan Hermann, and Steen Lassen, editors. 2004. *Dansk litteraturs kanon: rapport fra kanonudvalget*, 1. udg. ; 1. opl edition. Undervisningsministeriets forlag, København.
- Hertel Hans. 1983. *Den daglige bog: bøger, formidlere og læsere i Danmark gennem 500 år*. Forening for Boghaandværk, Kbh.
- George Levine. 2008. *How to read the Victorian novel. How to study literature*. Blackwell Pub, Malden, MA.
- György Lukács. 1964. *Probleme des Realismus*. Neuwied: Luchterhand.

- Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and Thamar Solorio. 2018. [Letting emotions flow: Success prediction by modeling the flow of emotions in books](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 259–265, New Orleans, Louisiana. Association for Computational Linguistics.
- Anne-Marie Mai. 2016. [Canons and Contemporary Danish Literature](#). *Folia Scandinavica Posnaniensia*, 19(1):109–132.
- Anne-Marie Mai. 2022. *Danish Literature from 1000 to 1900*. Syddansk Universitetsforlag/University Press of Southern Denmark, Odense.
- Lone Kølle Martinsen. 2012. [Bondefrihed og andre verdensbilleder idehistoriske studier af b.s. ingemanns danmarkshistorie 1824-1836](#). *Temp - tidsskrift for historie*, 3(5):75–103.
- Pascale Moreira and Yuri Bizzoni. 2023. [Dimensions of quality: Contrasting stylistic vs. semantic features for modelling literary quality in 9,000 novels](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 739–747, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Franco Moretti. 2000. Conjectures on World Literature. *New Left Review*, (1):54–68.
- Franco Moretti. 2007. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, London New York.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Munch-Petersen Erland. 1978. *Romanens århundrede: studier i den masselæste oversatte roman i Danmark 1800-1870*. Ph.D. thesis, Forum, Kbh. Book Title: Romanens århundrede : studier i den masselæste oversatte roman i Danmark 1800-1870. ISBN: 9788755307179.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Jack Douglas Porter. 2018. Popularity/prestige. Technical report, Stanford Literary Lab.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Ally Smid. 2022. In de nieuwe literaire canon moet je de schrijfsters met een lampje zoeken. *Trouw*.
- Enzo Terreau, Antoine Gourru, and Julien Velcin. 2024. [Capturing Style in Author and Document Representation](#). ArXiv:2407.13358 [cs].
- Willie Van Peer. 2008. [Ideology or aesthetic quality?](#) In Willie Van Peer, editor, *The Quality of Literature: Linguistic Studies in Literary Evaluation*, volume 4 of *Linguistic Approaches to Literature*, pages 17–29. John Benjamins Publishing Company, Amsterdam.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17:261–272.
- Robert von Hallberg. 1983. [Editor’s Introduction](#). *Critical Inquiry*, 10(1):iii–vi.
- Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. [Can Authorship Representation Learning Capture Stylistic Features?](#) *Transactions of the Association for Computational Linguistics*, 11:1416–1431. Place: Cambridge, MA Publisher: MIT Press.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. [Improving text embeddings with large language models](#).
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. [Multilingual e5 text embeddings: A technical report](#). *arXiv preprint arXiv:2402.05672*.
- Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. 2019. [Success in books: predicting book sales before publication](#). *EPJ Data Science*, 8(1):31.
- Ian P. Watt. 2001. *The rise of the novel*. University of California Press, Berkeley.
- Mary Ann Frese Witt. 2000. [Are the Canon Wars Over? Rethinking Great Books](#). *The Comparatist*, 24:57–63.
- Yara Wu. 2023. [Predicting the Unpredictable. Using Language Models to Assess Literary Quality](#). Master’s thesis, Uppsala University, Uppsala.
- Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. [Perplexing canon: A study on GPT-based perplexity of canonical and non-canonical literary works](#). In *Proceedings of the 8th Joint SIGHUM*

Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), pages 172–184, St. Julians, Malta. Association for Computational Linguistics.

A Distribution of titles

We see the distribution of titles per category in our corpus (1870-1900) in Figure 5.

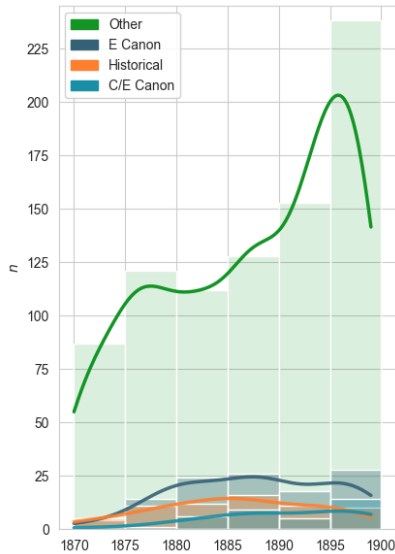


Figure 5: Distribution of titles per category in our corpus (1870-1900). This figure also reflects the actual incline of novels published in the period.

B Dendrograms

We see the dendrograms visualized in Figure 6.

C Intra and inter-group similarity

When calculating the similarity of books within groups over time (intra-group similarity), as well as the similarity between the canon and non-canon group over time (inter-group similarity), we used embeddings based on the m-e5-large-instruct model, using the prompt “Identify the author of a given passage from historical Danish fiction.”. The direction of change of both intra- and inter-group similarity proved consistent when using embeddings based on other models. For reference, in Table 3, we show the correlation over time for the various models tested. In the same table, we also show the correlation when using unprocessed means of rolling windows over time versus when using means of simulated distributions, as described in Section 5.2 ($s = 4, step = 1$). Note, again, that correlations vary slightly for each time we run the rolling window with simulated means (1,000 simulations of Gaussian distributions per window), so that the correlation coefficient should be taken as an indication rather than an exact value.

D Cosine similarity ranges

It is clear that cosine similarities are very high in our analysis. As noted, we use pooled embeddings which may affect a higher cosine similarity due to information loss. However, cosine similarity values are also high when comparing embeddings of raw chunks. As noted in the model card, the m-e5-large-instruct, cosine similarity scores of embeddings produced with this distribute in a narrow (and high) range¹⁹. Developers note: “This is a known and expected behavior as we use a low temperature 0.01 for InfoNCE contrastive loss. For text embedding tasks like text retrieval or semantic similarity, what matters is the relative order of the scores instead of the absolute values, so this should not be an issue.”

In Fig. 7, we show the distribution of cosine similarities for both raw and pooled embeddings for our corpus. Note that while the distribution of pooled embeddings does show a skew toward higher cosine similarity values, cosine scores of raw embedding chunks also exhibit a narrow range with a high mean. We therefore consider the very high cosine similarity scores an artefact of the model, rather than an effect of the pooling procedure per se.

E Canon directionality

We see the PCA with the mean embeddings of the canon/non-canon, plotted with a rolling window, visualized in Figure 8.

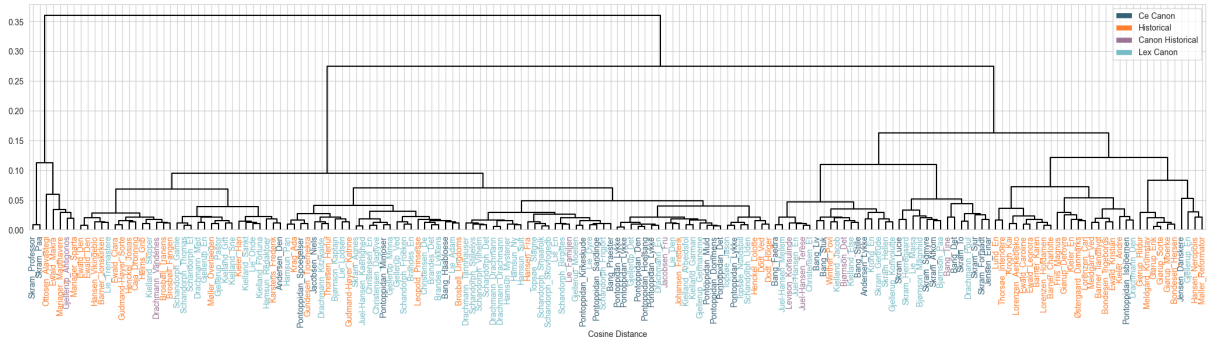
F Model References and Names

As many models are often updated, leading to a change in their output, we ensure reproducibility by specifying the revision IDs used in Table 4. The table also maps short-form model names used in the paper with their reference names as they appear on Huggingface.

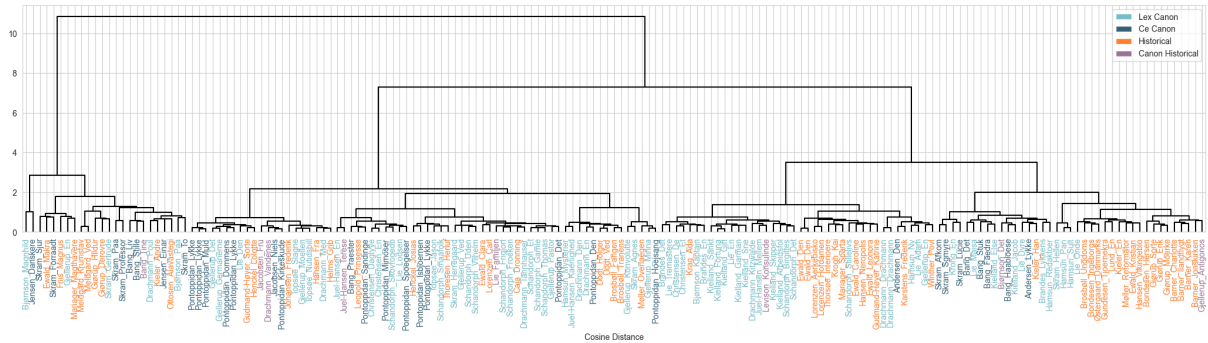
G Construction of Historic Evaluation task

Using the MiMe-MeMo corpus (Bjerring-Hansen et al., 2022), we similarly construct a clustering task as done in (Muennighoff et al., 2023; Enevoldsen et al., 2024). We down-sampled our corpus to 64 novels and took 32 chunks from each, adding up to 2,048 passages. The goal of the task is to see how well clusters of the embeddings correspond with the original authors. Clustering

¹⁹See FAQ, question 3, <https://huggingface.co/intfloat/multilingual-e5-large-instruct>



(a) embeddings



(b) TF/IDF values

Figure 6: Dendrograms based on cosine similarity of semantic embeddings (a) and TF/IDF (b). Dendrograms were calculated using Ward variance minimization, as implemented in SciPy v1.14.1 (Virtanen et al., 2020). Note that titles in purple are historical novels by – in our wider definition – canonical authors ($n = 171$).

	m-e5-large	m-e5-l-instruct (<i>identify</i>)	m-e5-l-instruct (<i>retrieve</i>)	DFM-large	MeMo-BERT
Canon	-0.61 (-0.82)	-0.48 (-0.62)	-0.59 (-0.73)	<u>-0.68</u> (-0.83)	-0.78 (-0.77)
Non-canon	<u>-0.81</u> (-0.87)	-0.87 (-0.90)	<u>-0.81</u> (-0.84)	-0.76 (-0.75)	-0.77 (-0.81)
Total	<u>-0.80</u> (-0.85)	-0.86 (-0.91)	-0.78 (-0.82)	-0.76 (-0.79)	-0.72 (-0.76)
Canon/non-canon	0.67 (0.80)	0.70 (0.74)	0.70 (0.77)	0.62 (0.84)	<u>0.68</u> (0.66)

Table 3: Correlation of intra and inter-group similarity over time using embeddings based on all models. Correlation over time based on the rolling windows’ simulated means and correlation over time between actual values *in parenthesis*. Note that we show the results of the m-e5-large-instruct model when instructed with two different prompts, “retrieve” and “identify”, see Table 4 in Appendix H for the full prompt, prompt 1 & 5. The strongest correlation is in bold, the second strongest is underlined. For all correlations, $p < 0.01$.

is performed using a K-means clustering of the authors of the passages. The performance is measured using V-scores similar to SEB (Enevoldsen et al., 2024). For the prompt-based model, we used the prompt “Identify the author of a given passage from historical Danish fiction”. To encourage future evaluations of historical Danish and Norwegian documents, we contribute our newly developed task to the Scandinavian Embedding Benchmark (SEB) in a pull request: <https://github.com/KennethEnevoldsen/scandinavian-embedding-benchmark/pull/184>.

H Instruction prompts

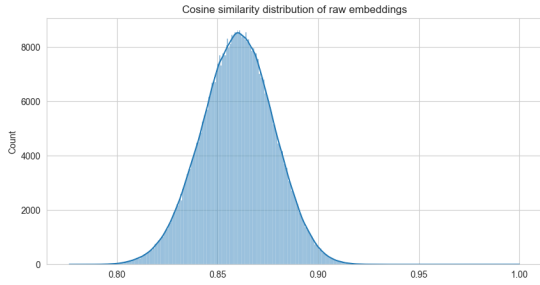
When generating the prompt, we followed the format used in (Wang et al., 2024a), where instructions for all clustering tasks start with the word “Identify”. We evaluated the performance of several versions of our final prompt on the custom historical task, which can be seen in Table 5.

Specifics of the formulation do not seem to have a large impact on performance; Prompts 1 and 2 perform similarly. Performance drops with prompts 3 and 4, which instructs the model to perform a different task, than it is evaluated on (cluster on books instead of authors). Finally, using a differ-

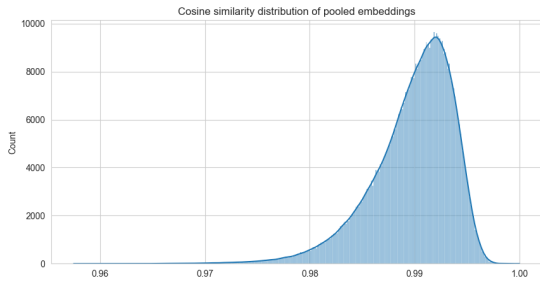
Name in Paper	Reference Name and Revision
m-e5-large-instruct	intfloat/multilingual-e5-large-instruct baa7be480a7de1539afce709c8f13f833a510e0a
m-e5-large	intfloat/multilingual-e5-large ab10c1a7f42e74530fe7ae5be82e6d4f11a719eb
DFM-large	KennethEnevoldsen/dfm-sentence-encoder-large-exp2-no-lang-align ec8293d8f447023de99d1e7fb79aa918d6258dc7
MeMo-BERT	MiMe-MeMo/MeMo-BERT-03 04cad875b848b56d9a76e80a031d60d66ae9cd02

Table 4: Model names as their appear in the paper along with reference name as their appear of hugging face along with the revision ID.

ent task keyword (e.g. "Classify" instead of "Identify") has some impact on the performance, as can be seen in prompts 5 and 6. This is likely the result of the training procedure of m-e5-large-instruct, as the model learns to embed the text conditional on the task prompt. For example, with a task definition that asks the model to retrieve, the model is likely trying to find a good query vector that lands close to relevant documents in embedding space, instead of embedding similar documents close together as is the goal of a clustering task.



(a) Cosine similarity range of raw embedding chunks. For this figure, we used the 20th chunk of each book and calculated cosine similarity between all pairs.



(b) Cosine similarity range of pooled embeddings of all books.

Figure 7: The range of cosine similarity scores for raw and pooled embeddings of the m-e5-large-instruct used for the main analysis, i.e., with prompt (1), see Table 5 in Appendix H.

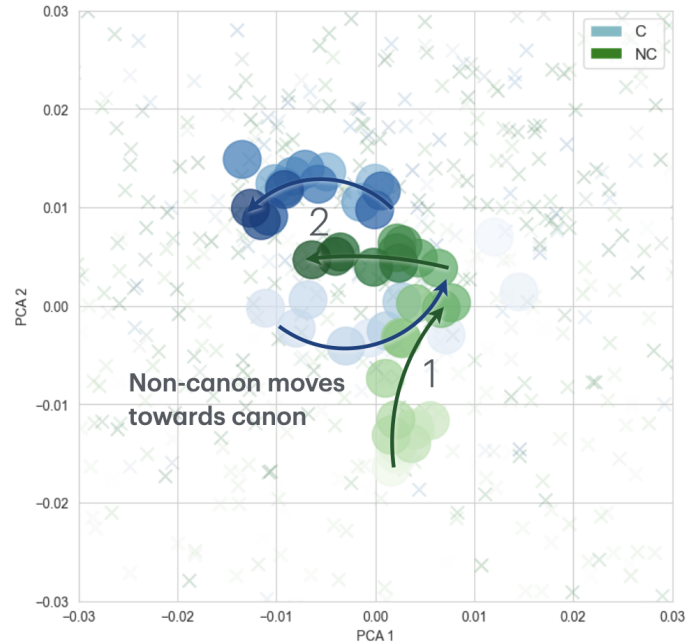


Figure 8: Positions of the mean embedding of the canon/non-canon, plotted with a rolling window size of 5 years, step 1. Time is indicated by the shading: darker colors are later in time. Note that the non-canon moves towards the canon. The PCA was fitted to all embeddings, then mean embeddings per group per window were fitted into the same PCA.

ID	Prompt	Historical
1	Identify the author of a given passage from historical Danish fiction	40.10
2	Identify the author of a specified passage taken from historical Danish literature	42.29
3	Identify which book from Danish historical fiction does the passage belong to	33.04
4	Identify the work from Danish historical fiction to which the provided passage belongs	34.35
5	Retrieve the author of a given passage from historical Danish fiction	42.56
6	Classify the author of a given passage from historical Danish fiction	46.43

Table 5: The performance of m-e5-large-instruct on SEB’s custom historical task using different prompts.

Deciphering psycho-social effects of Eating Disorder : Analysis of Reddit Posts using Large Language Model(LLM)s and Topic Modeling

Medini Chopra

Ashoka University / India
medini.chopra@alumni.ashoka.edu.in

Anindita Chatterjee

Peerless Hospitex Hospital
And Research Center Ltd. / Kolkata
anindita65@gmail.com

Lipika Dey

Ashoka University / India
lipika.dey@ashoka.edu.in

Partha Pratim Das

Ashoka University / India
ppd@ashoka.edu.in

Abstract

Eating disorders are a global health concern as they manifest in increasing numbers across all sections of society. Social network platforms have emerged as a dependable source of information about the disease, its effect, and its prevalence among different sections. This work lays the foundation for large-scale analysis of social media data using large language models (LLMs). We show that using LLMs can drastically reduce the time and resource requirements for garnering insights from large data repositories. With respect to ED, this work focuses on understanding its psychological impacts on both patients and those who live in their proximity. Social scientists can utilize the proposed approach to design more focused studies with better representative groups.

1 Introduction

Eating disorders (ED) are an area of increasing concern even as it continues to be under-reported and under-researched. Eating disorders are recognized by signs and symptoms which are published in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Clinical symptoms of ED most often manifest as Anorexia Nervosa, Bulimia Nervosa, Binge Eating Disorder, or ARFID which is defined as an eating or feeding disturbance so pervasive that the person is unable to meet appropriate nutritional needs, resulting in significant weight loss, nutritional deficiency, dependency on nutritional supplements, or interference in social functioning.

In (Silén and Keski-Rahkonen, 2022), it was reported that there is an alarming rise in ED across the world. According to the paper, an estimated 5.5–17.9% of young women and 0.6–2.4% of young men have experienced a DSM-5 eating disorder by early adulthood in the Western world. Studies from other parts of the world also show similar trends. A study conducted by NIMHANS in India in 2018

reported that Eating Disorders (ED) affect 6.5% of adolescent girls in India and an estimated 2–3% of the population as a whole. National Eating Disorders Association¹ reinforces this and states that though eating disorders impact people of all genders, ages, races, religions, ethnicities, sexual orientations, body shapes, and weights, the diversity is not represented in published research, which has most often relied on surveys of targeted groups. This lack of inclusion implies that the true impact of ED is still not well understood.

With the advent of hyper-specific online communities, those suffering from eating disorders (EDs) have access to social network platforms where they can share their experiences anonymously and seek support from each other as well as from therapists, counselors, and caregivers. Reddit, a popular online forum, hosts many such communities centered around eating disorders, with the largest one being r/EatingDisorders. These platforms are good sources of data since they are used by patients as well as those who live in proximity. Hence unlike targeted surveys which analyze information from affected individuals only, social media platforms also give insights into the psychological stress and trauma experienced by other groups of people like parents, partners, siblings, and friends of affected individuals.

In this paper, we present a study that exploits large language models (LLM) and topic modeling for analyzing large volumes of Reddit posts, to obtain insights about ED and its impact on social relationships. While traditional topic modeling techniques like Latent Dirichlet Allocation (LDA) perform a global analysis of large repositories, LLMs are good at generating summaries and answering questions through contextual analysis of individual articles, like posts. However, being gen-

¹<https://www.nationaleatingdisorders.org/resource-center/>

erative in nature, the answers to different questions have to be further semantically analyzed to identify global patterns. Most of the earlier work on analyzing social networks for similar purposes has been done on much smaller datasets. The research contributes to the growing body of literature on the intersection of mental health, social media, and Natural Language Processing. Since the posts are made available publicly after anonymization and careful removal of all sensitive information, these kinds of studies do not violate any ethical norms. It reinforces the potential of online platforms as a resource for studying Eating Disorders. It also establishes the role of LLMs in facilitating large-scale automated analysis in social science as stated in (Ziems et al., 2024).

The focus of our work is on analyzing the psychological state of people suffering from ED and those who live in proximity, using both statistical technical techniques like LDA and LLMs. Consequently, the key research questions were formulated as follows:

- Q1: What are topics that come up in the discussions on subreddit forums? Are there any new insights that can be obtained about the causes of ED and its effects on the physical and mental health of subjects? How do traditional methods like LDA compare with LLM-identified themes?
- Q2: Who are the common speakers in these posts, what are their relationships with people suffering from ED and what is the effect of this relationship on both? By doing so, we aim to understand not only what is being discussed, but how these discussions are framed and influence interpersonal dynamics.
- Q3: What could be a typical pipeline for analyzing large volumes of social media posts using LLM-generated insights? Besides giving directions on formulating the right prompts, we have also performed manual evaluations for parts of the content. This was done by identifying the original texts using linguistic phrases extracted, along with the LLM-generated inferences, evaluating the answers to questions on dominant emotions, and validating the mentioned interpersonal relationships, and their contexts from the text. From a computational point of view, this question

helps to establish the effectiveness of using LLMs for such research.

The rest of the paper is organized as follows. Section 2 presents an overview of earlier work in the area. Section 3 presents statistics about the data. Section 4 presents the analytical pipeline, along with details of each module. This is followed by results and discussions.

2 Related Work

Eating disorders (EDs) are complex psychiatric disorders with a range of associated mental and physical health symptoms. A meta-review article published in 2021 (Qian et al., 2021) analyzed thirty-three studies published across the world and concluded that the prevalence of eating disorders might have been underestimated thus far. Combined analyses showed that the lifetime prevalence of EDs in Western countries was 1.89%, and was high at 2.58% in females. Most of these studies did not use the DSM-5 criteria for their analyses. It also proposed that new diagnostic criteria should be used to comprehensively assess all types of eating disorders. In a study from Finland (Silén et al., 2020), the prevalence of lifetime DSM-5 eating disorders among the Finnish population was estimated to be 17.9% for females and 2.4% for males, and 10.5% across genders. A meta-review on ED-related research in India was presented in (Vaidyanathan et al., 2019), and also states that cultural context should be included in the analysis, with the help of culturally sensitive instruments used for diagnosis, and generating locally relevant epidemiological data about eating disorders from community as well as hospital settings.

The role played by social network platforms in the health and wellness pursuits of people was comprehensively analyzed in (Marks et al., 2020). The article states that social networking sites (SNSs) have evolved into an informal source of health education which influences peoples' health choices. Analysis revealed that the discussion majorly focuses on the promotion of idealized bodies, healthy foods, diets, and exercise. While this raises concern about the influence of social media on mental health, especially promoting body image concerns, eating disorders, and psychological distress, it is also acknowledged that social media can be alternately used to encourage wellness through promoting self-acceptance, intuitive eating, and life-enhancing movements. This work critically eval-

uates both the positive and negative role of social networks concerning health and wellness content. Prior to this, a small sample study on content analysis from a pro-ED community on Reddit was presented in (Sowles et al., 2018). This study delved into the task of identifying the expression of behaviors aligned with ED symptoms and support for these behaviors within the social network discussion. The paper presented results based on an analysis of four weeks of topic-specific discussion threads on a small dataset of 125 threads created for the purpose. It reported statistics on behaviors consistent with ED, and how the platforms provided support. Topic modeling using Latent Dirichlet Allocation (LDA) was employed to identify topics in eating disorder-specific social media content in (Moessner et al., 2018). The paper employed social network analysis using a linear network autocorrelation model to analyze communication patterns and the most influential users.

Twitter, another popular social network platform was studied by (Zhou et al., 2020) to assess social media engagement among individuals with ED. The focus of this was to develop an automatic approach to assess public perceptions about ED and ED-related behaviors. A machine learning approach was developed to identify ED-relevant tweets, and content analysis using topic modeling was implemented to unearth potential ED-related factors including behaviors, thoughts, and mental status.

3 Dataset Details and Data Preparation

We scraped data from three Reddit communities: r/EatingDisorders, r/EatingDisordersOver30, and r/EDAnonymous, spanning from 2020 to 2023, among which the most prolific one was r/EatingDisorders. A total of 20,918 posts and 58,228 comments posted from 2021 to 2023 were collected from this site. Upon generating a time-based chart with the number of monthly posts, we found a significant increase in the numbers over this period. The most substantial spike occurred between 2021 and 2022 as shown in Figure 1, and continued consistently into 2023. For this paper, we have presented insights extracted from the posts that appeared in the years 2022 and 2023, since these would present the most contemporary issues. The results presented are based on an analysis of more than 16.5 K posts, with 6700 from 2022 and 10450 posted in 2023. Text cleaning and pre-

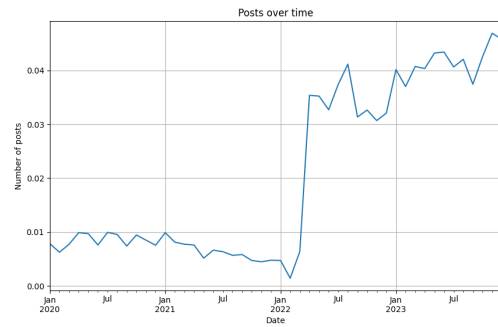


Figure 1: Number of Posts in subreddit community Eating Disorders

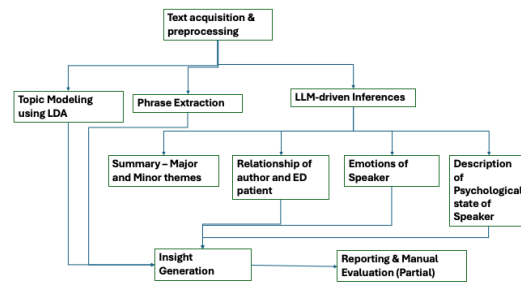


Figure 2: Unsupervised Content Processing Pipeline using traditional NLP techniques and LLMs

processing included the removal of HTML tags, URLs, special symbols, etc. The title and text were merged into one body for analysis.

4 Methodology

Figure 2 describes the processing pipeline, which is generic and can be applied for analysis of any large-scale repository. Starting with an aggregate analysis of content, we move towards an analysis of individual posts. The first step towards aggregate analysis is topic modeling. Statistical topic modeling techniques like Latent Dirichlet Allocation (LDA) leverage an unsupervised statistical approach to discover hidden semantic patterns from a large text collection to automatically identify topics that exist inside it. Topics are represented by clusters of similar words within a body of text.

Though topic modeling is effective in obtaining a bird’s eye view of content from a large repository, the words assigned to topics by themselves do not obtain much semantic information. To alleviate the problem, we have exploited the generative powers of an LLM to generate a semantically coherent description of the topic from the representation words. This step is referred to as Topic naming.

The topic summaries generated at this step can be used to compare topics against each other. This has been utilized later to compute topic trends based on the percentage presence of similar topics across the years.

The aggregate analysis is followed by an individual analysis of posts. Since the intent of analyzing any repository is to obtain statistics about certain key concepts present in the repository, the objective of this step is to establish the presence or absence of the concepts in each post. This step is akin to "coding" in the social sciences, which is an analytical process by which data elements are "categorized" to facilitate analysis. Traditionally this is an expensive knowledge-intensive task. In recent times, there have been efforts to use the Large Language Models (LLMs) for this task, as these are capable of performing various language processing tasks without additional task-specific training. These models, termed as zero-shot LLMs, can also reliably classify and explain certain social phenomena (Ziems et al., 2024). It is this aspect of a large language model, that we have exploited to use it as a computational social science tool. Details for each of these sub-tasks are presented in the following subsections.

4.1 Obtaining Topical Insights from Post repositories

To explore the nature of conversations and the content of discourse in r/EatingDisorders, we implemented Latent Dirichlet Allocation (LDA) for topic modeling. This technique identifies hidden topics in text data by grouping words that frequently occur together. We conducted separate topic modeling for 2022 and 2023 data, to obtain insights about the temporality of topics, including the emergence, persistence, and disappearance of topics over this time. The optimal number of topics was decided based on coherence, inclusivity, and perplexity scores, and was found to be 16 for each year. On average, we achieved a perplexity score of -22 and a coherence score of 0.40, indicating a reasonably well-fitted model.

LDA assigns a topic distribution for each of the 16 topics to each post, where the total topic strength sums up to 1. Each post may contain a significant presence of more than one topic. To determine how good the topic assignments are, we performed skew analysis for the entire sets of each year. Skewness is a statistical measure that reveals whether a distribution is symmetric or asymmetric. For asymmetric

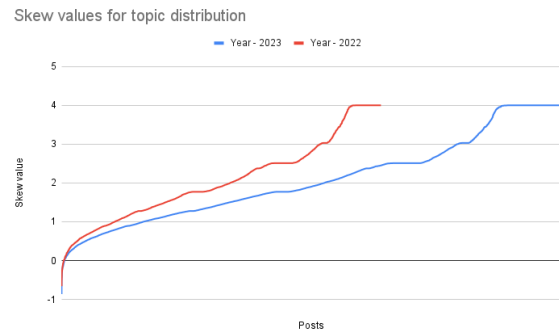


Figure 3: Skew Values for topic Distribution: Positive values indicate the presence of a dominant topic

distributions, the skew values reveal whether there is a sharp trend towards either end of the normal curve. For topic assignments, a zero skew value would indicate that all topics are present in equal measure thereby revealing the absence of any significant topic in the text. A positive skew value on the other end would indicate the presence of dominant topics. Figure 3 shows that the skew values are majorly positive, indicating that the topic distributions can be used to find the dominant topics of the posts. To analyze the topics and their relative presence in the discussions, each post was assigned a "dominant topic" which was the topic with the maximum strength for it. The red line is shorter because 2022 had fewer posts than 2023.

For each topic identified, LDA returns the probability distribution of the words to the topics. Using the top 30 words for each topic, a one-line description and a short name were generated for it, with the help of gpt-3.5-turbo-1106. The summaries for each topic thus obtained were compared with each other within and across the years to identify similar topics, if any. BERT-based embeddings were generated for each summary, and then cosine similarity measures were computed between the topic summaries identified for 2022 and 2023. To compute the topic strength for a year, we computed the percentage of posts that had the corresponding topic as the dominant topic in that year. This analysis allowed us to plot the relative strengths of topics each year, revealing shifts in focus across the two years. The results were further contextualized through consultations with a psychologist to understand the emotional and psychological dynamics of these communities.

4.2 Analysis of Posts using Large Language Model

After obtaining the aggregate views on topics in the posts, analysis of individual posts was done by employing prompt engineering to generate structured summaries of posts. The process was designed in a way that helps distill complex discussions into manageable insights. First, a set of intent was created, based on the analytical insights required. For each intent thereafter, a prompt was designed to extract relevant information from the post. The intent and the corresponding prompts used for this work are as follows:

- Intent - Summarize
Prompt 1: Write a 1 sentence summary of the post text and identify the major and secondary themes
- Intent - Identify emotions expressed in post
Prompt 2: What are the major emotions conveyed in the post?
- Intent - Establish a relationship between author and subject suffering from ED
Prompt 3: Identify whether the author of the post is writing about their own experience or discussing the issues concerning someone else. What is the relationship between the author of the post and the person whose ED-related disorders are discussed?
- Intent - Understand the psychological impact of ED
Prompt 4: What can you infer about the impact of ED on the psychological state of the author?
- Intent - Identify the presence of DSM 5 Eating Disorder symptoms in post
Prompt 5: What are the physical, physiological, or psychological impacts of ED mentioned in the text?

We used the OpenAI gpt-3.5-turbo-1106 model, with the parameters temperature set to 0 and maxtokens set to 150. Temperature is used to control the randomness in the response generated; when the temperature is low, it chooses only the most likely next word based on the context. We kept it at 0 to ensure minimum randomness in the responses generated. The *maxtoken* parameter defines the limit on the number of tokens the model can generate. The prompts were executed

with each post given as the context for generating the answers. For clarity of presentation, we have segregated the prompts, though they were all executed together. Prompt 5 could be broken into more atomic-level prompts, one for each disorder. However, it would have incurred more computational cost. There were further instructions to restrict the prompts to generate answers in very few words, and repeat words for similar interpretation. Though posing multiple choice-type questions could be a way to alleviate the problem, we refrained from doing so since there are reports that LLMs assign the choices quite randomly, rather than logically, whereby the answers on the top are returned most often. In contrast, the generative approach allows it to generate a more plausible answer (Li et al., 2024).

It was observed that though there was substantial repetition in the answers, there were some syntactic variations within semantically similar answers. For example, for Prompt 2, while the most common answer was "the speaker is referring to themselves", similar answers were "the speaker is likely referring to themselves", "self-referential" etc. In case the author is not talking about self, common answers were "the speaker is in a romantic relationship with the person being referred to" or "the speaker is the parent of the child they are referring to" etc. Some additional effort was required to aggregate the answers under different categories, for which string-matching codes were written.

5 Results: Findings and Discussion

We first present an aggregated analysis of the content obtained in terms of topics and topic trends using LDA. A short description of each topic was generated by feeding the top 10 topic words to LLM. To compare the topic strengths for two years, topic alignment was done using BERT-based similarity scores among the short descriptions across the years. The unique topics, their names, and trends across the years found by LDA are presented in Figure 4. While the most dominant topic of 2022 was "weight gain and image-related issues", the conversation around this substantially dropped in 2023. The topics rather shift towards those which are about "seeking help", sharing experiences on "relapse and recovery", "disordered eating", and "eating patterns and appetite" along with a discussion on handling "relationship issues" and advice on handling "social interactions". There are also

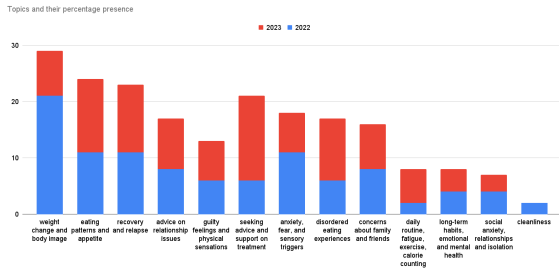


Figure 4: Topic Presence in percentage of posts for 2022 and 2023

topics centered around extreme feelings of guilt, shame, concern, and fear which appear as a major topic. These emotions will be discussed in more detail later. A topic centering around "cleanliness" discussing changes in personal cleanliness routines and organizational approaches to cleanliness featured in 2022, but not in 2023. This could be due to the immediate aftermath of the pandemic. The changes in topic trends indicate that the platform is indeed emerging as a forum where ED patients form a mutual support group.

Figure 5 presents the most frequently occurring topics in the posts as identified by the LLM using prompt 1. Comparing these with figure 4 shows that while both the results show similar trends, the LLM-identified topics are broader in nature, more like themes rather than topics. An issue faced with LLM themes was the wording differences in the answers generated for the posts. These themes had to be combined based on the words that appeared in them.

The similarity of the topics and themes obtained by two different unsupervised methods increases confidence in the results obtained. The most commonly occurring theme is that of personal struggle faced by patients of ED, especially during recovery, as they frequently get into relapse. This is followed by concerns about body image which is at the center of triggering ED in people. Impact on mental and physical health is also a recurring theme. The topics reveal finer nuances of these topics as weight-related issues for body image, guilty feelings along with anger, fear, and anxiety as chief mental health problems.

Few posts had explicit mentions of the age and gender of the author or the patient of ED as "x/G" where x indicated age and G indicated gender. 208 posts in 2022 and 351 posts in 2023 had such mentions. Figure 6 shows their relative distri-

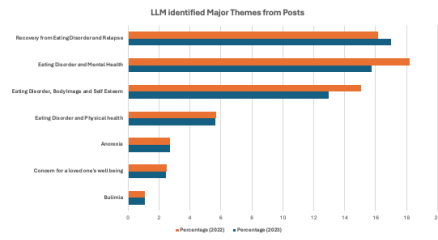


Figure 5: Post Themes identified by LLM and their percentage presence in 2022 and 2023

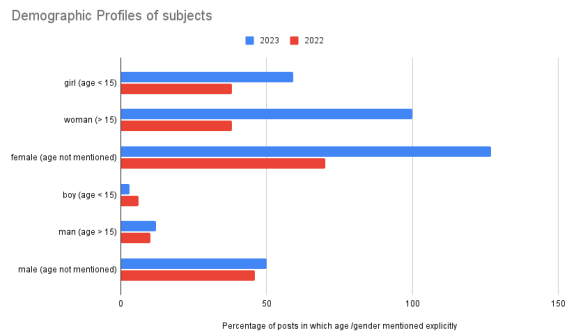


Figure 6: Demographic Profiles of users who explicitly mention age, gender - indicates trends and not estimates

butions. There was a much higher mention of woman/female/girl than man/male/boy across both years. This is consistent with the findings in past literature where women are reported as more susceptible to eating disorders than men. However, this also shows the spread or acknowledgment of ED across demographics. These results establish that monitoring social media platforms could provide a good understanding of trends and thereby help clinical experts choose their subjects in a more informed manner for deeper studies. We also found that a significantly higher number of mentions came from the teenage group, peaked at around 25, and then started decreasing, with the higher mentions of age, going up to 50+. The lowest age mentioned is 10, which also aligns with the literature. All results were manually verified and found to be correct.

Emotion analysis results were obtained from the output of Prompt 2. Analysis shows that the emotions present in the posts are predominantly negative, with concern, frustration, confusion, fear, and desperation dominating over others. Figure 7 shows that the relative presence of these emotions is fairly consistent over time. Such a high presence of negative emotions is consistent with the DSM 5

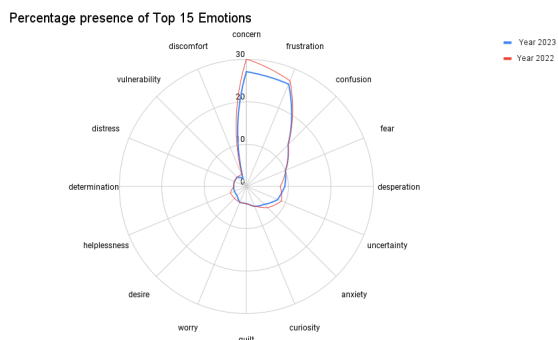


Figure 7: Emotion distribution across posts for 2022 and 2023

diagnostic disorders which mention that individuals who develop eating disorders are at increased risk of developing various co-occurring mental health concerns, including anxiety, depression, obsessive-compulsive disorder (OCD), and Post-traumatic stress disorder (PTSD).

A manual evaluation of the LLM-assigned emotion was done for 100 posts, 50 from among those tagged with *frustration* and another 50 tagged with *guilt and shame*. The evaluation was done by a psychology expert, who had no prior exposure to the data, but was explained about the process and informed that the tags were assigned by a language modeling tool. For the first set on *frustration*, the accuracy of LLM tagging was found to be 90%, for which the expert agreed that the post exhibited frustration. For the remaining posts, the expert marked the emotions as *confusion, obsession, etc.* and not frustration. For the posts marked with *guilt*, the expert agreed fully with the LLM tagging. On inspection of the posts, it was found that all these posts had explicit mentions of *guilty, guilt, shame*. Evaluations at a larger scale are being planned. However, the preliminary results do establish the effectiveness of using LLMs in large-scale data analysis for mental health from social networking sites. This also establishes the possibility of deploying LLMs to set up semi-supervised and distant-supervised learning frameworks, by providing a quick approach to generate training data, which is reasonably clean, and not expensive to initiate.

Figure 8 presents an analysis of answers generated in response to Prompt 3. It extracts the relationship of the authors of a post to its subject who is mentioned in it as the patient suffering from ED. These results are from the 2023 collection. As expected, around 70% of the authors are the patients

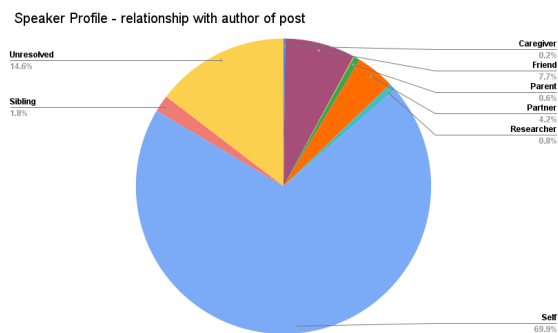


Figure 8: LLM-inferred relationship between author of post and subject of ED

themselves. However, there are a substantial number of posts that are written by friends and acquaintances, parents, partners, and siblings. Friends and acquaintances include colleagues, roommates, family friends, etc. Siblings are mentioned either as such or as brothers or sisters. The partner tag was assigned to authors who were identified by the LLM as "one in a romantic relationship with the person suffering from ED" or a spouse, girlfriend, boyfriend, or "someone in a close relationship with the patient".

Responses to Prompt 4 provide deeper insights into the psychological impact of ED on the patient when the author is "Self", as well as on other people when the author of the post is a friend, parent, partner, colleague, or sibling. These results are presented below.

5.1 LLM inferred psychological impact from posts written by others

- Speaker feels *anxious* (or *concerned, worried, frustrated, stressed*) by the state of the person suffering from ED: 67%
- Speaker is *unsure* (or *confused*) about how to help sufferer and is seeking help from the community as author: 37%
- Speaker feels *overwhelmed* by situation: 10%
- Speaker is *hopeful* of providing support to a person suffering from ED and seeks advice on how to do so: 15%
- Speaker is *worried* that their concern may hurt or induce feelings of embarrassment or shame or guilt in the person suffering from ED: 8%
- Speaker feels *triggered, spiraling, and struggling* with their mental health (indicating the

effect of proximity of ED patients on each other): 5%.

- Speaker *fears relapse* of ED in them due to the other person's diet-related disorders: 1%

Posts related to the above-mentioned situations highlight the difficulties faced by people who stay in proximity to people suffering from ED and hence are indirectly affected by it. It is evident from the representative posts that we have analyzed, that the conditions can pose serious health risks to many of them, as ED impacts their relationships. By talking about relapse, there is an indication that ED patients are affected by mutual proximity.

5.2 LLM inferred psychological impact of authors who are subjects of Eating Disorder

Overwhelmed: It is the most common inference drawn by LLMs to describe speakers' mental states. Summarily, speakers are overwhelmed by their mental state, managing weight-related issues and social pressures around it, relationship with food, and physical well-being - 15%. On analyzing the corresponding posts, these are found to be predominantly about user **struggles** during *recovery*, where they mention *stressing about weight issues, body image, self-perception*. These are directly mentioned in at least 26% of the posts, which in terms of absolute count is around 1800 posts.

Frustrated: Frustration surfaces whenever the user expresses an inability to handle their health, well-being, desires, cravings, or relationships. This shows up explicitly in 10% of the posts. Pain, functional dyspepsia, abnormal hunger, hypermetabolism, abnormal menstrual cycle, digestive disorder, and laxative abuse are some of the symptoms mentioned. It often indicates that the speaker feels misunderstood and unsupported, leading to increased frustration and likely exacerbating their struggle with the eating disorder.

Guilty: The speaker feels pressure and anxiety related to their eating and exercise habits, indicating a potential negative impact on their psychological state - 10%. Around 7.5% or 510 posts expressed explicit feelings of guilt due to the way the subjects feel about their bodies and hence became a subject of eating disorders. This is a complex feeling, where a feeling leads to an action, which in turn has inflicted remorse in the minds of the subject.

Family pressures and relationships: Around 7% of the posts contain discussions centered around

the role played by the family. A portion of these reports that *body shaming by family members* was either the cause of disorder or interfered with the recovery process. Another significant portion of these are about *how eating disorders of other family members are affecting them*. This indicates that proximity to a person with an eating disorder or obsession with body image induces an eating disorder. This points out the possibility of having *hubs of eating disorder* within spaces like hostels, schools, campuses, offices, or any other closed social neighborhoods where people eat together.

Help, Support, and Empathy: Speakers appeal for help, and seek support and empathy from other members to aid their recovery - 10%.

Binge eating: Speaker is experiencing anxiety and a sense of helplessness in their struggle with binge eating disorder - 2%. They have mentioned abnormal hunger, cravings for food, and inability to control their negative relationship with food.

Suicidal thoughts: Around 100 posts mentioned suicidal thoughts. This is also a known effect of ED. A low precedence of it in the platform, along with high volumes of appeal for help, reiterates the importance of social network platforms which act as support groups for patients, without feeling judged or misunderstood.

Academic Pressure: Around 40 posts mention academic struggles as their stress inducers.

Ethnicity: Though extremely few, the mention of *Asian* and *African* in posts indicated family-related struggles due to different cultural contexts at home including food, body shapes, and eating habits.

These insights can act as a springboard for designing more focused surveys for targeted groups with better representation.

6 Conclusion

In this paper, we have presented a large-scale study on Reddit posts related to Eating Disorder. We have shown how large language models like GPT can be used effectively to gain insights about facts and figures using their power to summarize and draw inferences against targeted prompts. The most significant insight gained through this study is about people who live in proximity to the subject of Eating disorder, and use the platform to convey their distress. The psychological stress undergone by this group which includes parents, siblings, friends, and partners is not much reported. Though the known effects of ED are found in the data, the

study reveals that the platforms are primarily used for advice, support, and seeking empathy from co-users. The results are highly encouraging. Our plan includes devising a way to methodically evaluate the LLM-generated answers. The ultimate intent is to build a framework for analysis of mental health-related posts, that can expedite the whole process of insight generation as well as strategically plan interventions and support mechanisms.

7 Limitations

The limitation of the present work lies in the fact that only a small number of evaluations could be carried out. This is due to a lack of funding and expert resources.

Acknowledgments

This research was supported by the Ashoka Mphasis Lab - a collaboration between Ashoka University and Mphasis Limited.

References

- Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. [Can multiple-choice questions really be useful in detecting the abilities of llms?](#) *arXiv preprint arXiv:2403.17752*.
- Rosie Jean Marks, Alexander De Foe, and James Collett. 2020. [The pursuit of wellness: Social media, body image and eating disorders.](#) *Children and Youth Services Review*, 119:105659.
- Markus Moessner, Johannes Feldhege, Markus Wolf, and Stephanie Bauer. 2018. [Analyzing big data in social media: Text and network analyses of an eating disorder forum.](#) *International Journal of Eating Disorders*, 51(7):656–667.
- Jie Qian, Ying Wu, Fanxiao Liu, Yikang Zhu, Hua Jin, Hongmei Zhang, Yumei Wan, Chunbo Li, and Dehua Yu. 2021. [An update on the prevalence of eating disorders in the general population: a systematic review and meta-analysis.](#) *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, pages 1–14.
- Yasmina Silén and Anna Keski-Rahkonen. 2022. [Worldwide prevalence of dsm-5 eating disorders among young people.](#) *Current Opinion in Psychiatry*, 35(6):362–371.
- Yasmina Silén, Pyy N Sipilä, Anu Raevuori, Linda Mustelin, Mauri Marttunen, Jaakko Kaprio, and Anna Keski-Rahkonen. 2020. [Dsm-5 eating disorders among adolescents and young adults in finland: A public health concern.](#) *International Journal of Eating Disorders*, 53(5):790–801.
- Shaina J Sowles, Monique McLeary, Allison Optican, Elizabeth Cahn, Melissa J Krauss, Ellen E Fitzsimmons-Craft, Denise E Wilfley, and Patricia A Cavazos-Rehg. 2018. [A content analysis of an online pro-eating disorder community on reddit.](#) *Body image*, 24:137–144.
- Sivapriya Vaidyanathan, Pooja Patnaik Kuppili, and Vikas Menon. 2019. [Eating disorders: An overview of indian research.](#) *Indian journal of psychological medicine*, 41(4):311–317.
- Sicheng Zhou, Yunpeng Zhao, Jiang Bian, Ann F Haynos, Rui Zhang, et al. 2020. [Exploring eating disorder topics on twitter: Machine learning approach.](#) *JMIR Medical Informatics*, 8(10):e18273.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, arXiv:2305.03514(1):237–291.

Topic-Aware Causal Intervention for Counterfactual Detection

Thong Nguyen

National University of Singapore
e0998147@u.nus.edu

Truc-My Nguyen

Ho Chi Minh city University of Technology
nguyenmy3399@gmail.com

Abstract

Counterfactual statements, which describe events that did not or cannot take place, are beneficial to numerous NLP applications. Hence, we consider the problem of counterfactual detection (CFD) and seek to enhance the CFD models. Previous models are reliant on clue phrases to predict counterfactuality, so they suffer from significant performance drop when clue phrase hints do not exist during testing. Moreover, these models tend to predict non-counterfactuals over counterfactuals. To address these issues, we propose to integrate neural topic model into the CFD model to capture the global semantics of the input statement. We continue to causally intervene the hidden representations of the CFD model to balance the effect of the class labels. Extensive experiments show that our approach outperforms previous state-of-the-art CFD and bias-resolving methods in both the CFD and other bias-sensitive tasks.

1 Introduction

Counterfactual statements describe an event that may not, did not, or cannot occur, and the consequence(s) that did not occur as well (O’Neill et al., 2021). For example, consider the statement — *I would purchase this physics book, but I really want that my brain has a tiny amount of interest in science!*. We can partition the statement into two components: a component about the event (*my brain has a tiny amount of interest in science*) as the antecedent, and the consequence of the event (*I would purchase this physics book*) as the consequent. Both the antecedent and the consequent did not take place (*neither the speaker has purchased the book nor he is interested in science*). Accurate detection of such counterfactual statements is beneficial to various NLP applications, such as in social media or psychology. In social media, counterfactual detection (CFD) can be helpful by eliminating irrelevant content (O’Neill et al., 2021).

Scenarios	Examples	mBERT Predictions	Labels
Clue phrase Anomaly	It doesn’t work as well as I was hoping it would, it is a waste of money.	Negative	Positive
	I don’t like to go into the plot a lot. The blurb represents the book fairly.	Negative	
	Who would have thought a pillow could make such a difference.	Positive	Negative
	The girlfriend was annoying, and it made me wonder if any man in his right mind would have put up with her behavior as long as he did.	Positive	
Cross-lingual input	It would have been, people would say, worse than Watergate.	Positive	Positive
	ウォーターゲート事件よりもひどかったかもしれない、と人々は言うだろう。	Negative	

Table 1: Examples of counterfactual detection from the Amazon-2021 dataset. We denote mBERT predictions of *positive* (counterfactual) and *negative* (non-counterfactual) classes.

For example, in the previous statement, we should not return science or physics content because the user is not interested. Detecting counterfactuality can also give useful features to perform psychology assessment of huge populations (Son et al., 2017).

Previous development of monolingual and multilingual CFD methods depend on extensive labelled datasets (O’Neill et al., 2021). However, in CFD datasets, the percentage of counterfactual examples is heavily low, even approaching 1 – 2% (Son et al., 2017). This class imbalance has two weaknesses. First, because counterfactual hints are so limited for the CFD model to learn, it tends to rely on clue phrases, e.g. *if*, *I wish*, etc., to detect counterfactuality. When the existence of such clue phrases does not correlate with the counterfactuality, the model might be led to false predictions. As illustrated in Table 1, the mBERT baseline predicts all incorrect classes for both counterfactual examples, which do not include clue phrases, and non-counterfactual ones, which include clue phrases. Moreover, the performance might substantially drop if the model is tested upon a language different from the training language. As shown in Table 1, the multilin-

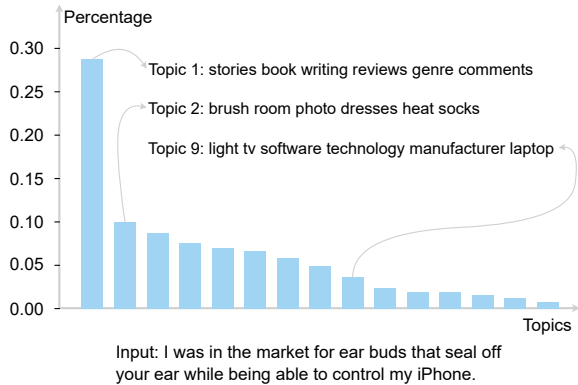


Figure 1: For each topic, we count the percentage of inputs in which the topic has the largest probability in the topic representation. Topic 1, 2, and 9 refer to three top topics of the input document, in descending order of probability.

gual mBERT predicts the correct class for the English statement, but misclassifies the Japanese one of similar meaning. Second, the class imbalance causes the CFD model to bias towards the non-counterfactual class over the counterfactual one, thus resulting in sub-optimal performance.

To address the first issue, we propose to incorporate neural topic model (NTM) into the CFD module. Particularly, we aim to approximate the global semantics of the input statement learned from the posterior distribution of the NTM. The posterior distribution generates the global semantics in terms of the topic representation to guide the CFD model towards semantics of the input instead of the clue phrases. However, a challenge exists that the NTM tends to repetitively assign large weights to a certain small group of topics. In Figure 1, even though the input statement is about *ear buds* and *iPhone*, the NTM still infers it to be highly related to *stories*, *book*, and *reviews*. To cope with this challenge, we propose to adapt backdoor adjustment that adjusts the behavior of neural topic model to make it consider all topics fairly. To the best of our knowledge, no study has explored the benefit of backdoor adjusted NTM for counterfactual detection.

To address the second issue, we view the CFD problem from a causal perspective. Our perspective gives rise to a causal graph where the class imbalance plays a confounder role in influencing hidden representations of the input statement. Based on the graph, we propose to perform causal intervention on these representations to remove the confounding effect of the imbalance phenomenon and enhance the model prediction.

To sum up, our contributions are as follows:

- We propose a novel neural topic model equipped with the backdoor adjustment to produce effective topic representations for benefiting counterfactual detection.
- We propose causal intervention upon hidden representations to ameliorate the confounding effect of the class imbalance in counterfactual detection datasets.
- Extensive experiments demonstrate that we significantly outperform state-of-the-art CFD and bias-resolving approaches. Our method is also applicable to other bias-sensitive natural language understanding tasks.

2 Related Work

2.1 Representational Intervention for Deep Learning

Representational intervention has been popularly adopted in deep learning applications. Some include document summarization (Nguyen et al., 2021; Nguyen and Luu, 2022), topic modeling (Wu et al., 2024c, 2023a, 2024d, 2023b, 2024a), document ranking (Nguyen et al., 2023c, 2022), sentiment analysis (Nguyen et al., 2023b, 2024a, 2023a), video moment retrieval (Nguyen et al., 2023d, 2024d), and video question answering (Nguyen et al., 2024c,e). As one notable approach for representational intervention, causal inference has attracted myriad attention as a method to interpret adversarial attacks (Zhao et al., 2022) and eradicate spurious confounding factors in SGD optimizer (Tang et al., 2020).

2.2 Predictive Biases in Deep Learning

Research community has long searched for objective-based and augmentation-based countermeasures against biases that drive deep learning models to ignore the input content when making predictions (Wu et al., 2024b; Nguyen et al., 2024b; Nguyen and Luu, 2021; Nguyen et al., 2024f). For the objective-based direction, Karimi Mahabadi et al. (2020) propose to increase the loss weight of rare examples and subtract the gradients of the biased model from the main one to mitigate their spurious influences. In the second direction, Wang et al. (2022) perturb words to prevent the confounding effect of language bias. Wang and Culotta (2021) suggest augmenting the original training

set with samples containing antonyms of high coefficient terms and reverse label. However, their method demands human supervision and solely involves sentiment classification. Focusing on Counterfactual Detection, O’Neill et al. (2021) decide to mask clue phrases and populate counterfactual examples through backtranslation. Nevertheless, they find that these methods suffer from deficiency since counterfactuality also depends on the context. Contrast to them, we decide to causally intervene the contextualized representations to reduce the confounding effect of the biases.

3 Methodology

In this section, we sequentially formulate the preliminaries of counterfactual detection and neural topic model, introduce our proposed causal perspective for the task, and then articulate the implementation details of our framework.

3.1 Problem Formulation

Given an input sentence $S = \{w_1, w_2, \dots, w_N\}$ and its bag-of-words (BOW) representation \mathbf{x}_{bow} , we aim to train a model function f that maps S and \mathbf{x}_{bow} to a probability scalar $y \in [0, 1]$. The probability magnitude will denote whether the input sentence is counterfactual or not.

3.2 Neural Topic Model (NTM)

Our neural topic model possesses the Variational AutoEncoder architecture (Miao et al., 2017; Kingma and Welling, 2013). It consists of an encoder to produce topic representation and a decoder to reconstruct the original input based upon the representation.

Topic Encoder. Its function is to encode the input \mathbf{x}_{bow} into the topic representation θ . In the beginning, \mathbf{x}_{bow} is forwarded to both non-linear and linear layers to estimate the mean μ and standard deviation σ of the variational distribution $q(\mathbf{z}|\mathbf{x})$:

$$\pi = f_0(\mathbf{x}_{\text{bow}}), \mu = f_\mu(\pi), \log \sigma = f_\sigma(\pi), \quad (1)$$

where we implement f_0 as a non-linear layer with the softplus activation function; f_μ and f_σ are two linear layers. Subsequently, to lessen the gradient variance, we adapt the reparameterization trick (Kingma and Welling, 2013) to draw the latent vector \mathbf{z} :

$$\mathbf{z} = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (2)$$

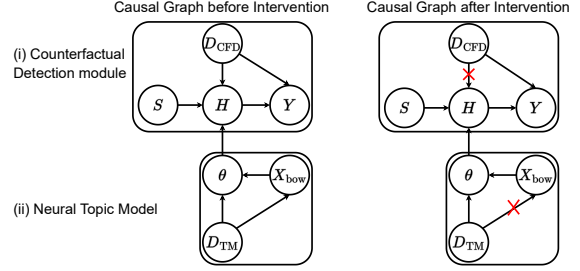


Figure 2: (left) Our proposed causal model for counterfactual detection. (right) The causal graph after removing arrows from D_{CFD} to H and D_{TM} to X_{bow} , eliminating spurious effects of the label and topic biases.

Then, we normalize \mathbf{z} with the softmax function to attain the topic representation θ as:

$$\theta = \text{softmax}(\mathbf{z}). \quad (3)$$

Topic Decoder. Given the topic representation θ , the decoder’s task is to reconstruct the original input \mathbf{x}_{bow} as \mathbf{x}'_{bow} . It performs the sampling process to extract the word distribution:

- For each word $w \in \mathbf{x}_{\text{bow}}$, draw $w \sim \text{softmax}(f_\phi(\theta))$,

where f_ϕ denotes a ReLU-activated non-linear transformation. In the ensuing sections, we designate the weight matrix $\phi = (\phi_1, \phi_2, \dots, \phi_K) \in \mathbb{R}^{V \times K}$ of f_ϕ as the topic-word distribution, in which V and K denote the vocabulary size and the number of topics, respectively. We also leverage the topic representation θ as global semantics to enhance the counterfactual detection module.

3.3 Causal Perspective into Counterfactual Detection

To investigate the relation among factors in our counterfactual detection system, we propose a structural causal graph (SCG) in Figure 2. Our graph includes vertices, direct edges, and two sub-graphs for denoting random variables, causal effects, with the pre-intervened and post-intervened states, respectively.

SCG for Topic Modeling. In this component, the topic bias D_{TM} is the confounder that influences variables θ and X_{bow} in the neural topic model.

- $X_{\text{bow}} \leftarrow D_{\text{TM}} \rightarrow \theta$: This backdoor path elicits the spurious correlation between \mathbf{x}_{bow} and θ instances. In topic modeling, neural topic models have a tendency to align documents with a repetitive set of topics.

- $\theta \rightarrow H$: Because of the confounder D_{TM} , the inferred global semantics might comprise irrelevant entries that do not represent the true semantics of the document. Therefore, the fallacious semantics could become detrimental noise to adulterate the hidden representations \mathbf{h} , which is the direct input to the counterfactual classifier.

SCG for Counterfactual Detection. This component delineates causalities among four variables in counterfactual detection: input sequence S , encoded content H , output prediction Y , and the imbalanced label bias D_{CFD} . In detail, the imbalanced label distribution confounds both the predicted output Y and variable H , leading to erroneous correlation between H and Y .

- $H \leftarrow D_{\text{CFD}} \rightarrow Y$ specifies the effect of D_{CFD} on hidden representations. In practice, the overwhelming population of the negative label in counterfactual datasets might result in learned representations that mostly express non-counterfactual features, thus driving the detection model towards the non-counterfactual response during prediction.

Causal Intervention on Textual Representations. We now present the method to remove the confounding effects. To obtain the deconfounded representations, we capture the causal effect from X_{bow} to θ and from H to Y via the Causal Intervention technique, i.e. Backdoor Adjustment (Pearl, 2009), with the following theorem to remove the arrow from D_{TM} to X_{bow} and D_{CFD} to H .

Theorem 1. (Backdoor Adjustment (Pearl, 2009)) Let $o \in \{y, \theta\}$, $i \in \{\mathbf{x}_{\text{bow}}, \mathbf{h}\}$, and $n \in \{d_{\text{TM}}, d_{\text{CFD}}\}$. Then,

$$p(o|\text{do}(i)) = p^{N \rightarrow I}(o|\text{do}(i)) = \sum_n p(o|i, n) \cdot p(n). \quad (4)$$

This theorem shows that we can model the deconfounded likelihood $p(o|\text{do}(i))$ through estimating $p^{N \rightarrow I}(o|i, n)$ and $p(n)$. We will expound the implementation of $p(o|\text{do}(i))$ to assist the model in predicting counterfactuality in Section 3.4 and deconfound neural topic model in Section 3.5.

3.4 Model Implementation

Our overall framework is illustrated in Figure 3.

NTM for Text Encoder. To address the issue of model reliance on clue phrases in counterfactual

detection, we propose to condition contextualized representations on global semantics yielded by the neural topic model.

Initially, we append the special token [CLS] to the beginning of the input sequence. Then, the text encoder converts each discrete token w_i into the hidden vector \mathbf{h}_i as:

$$\mathbf{h}_{[\text{CLS}]}, \mathbf{h}_1, \dots, \mathbf{h}_{|S|} = \text{TextEncoder}([\text{CLS}], w_1, \dots, w_{|S|}). \quad (5)$$

We insert global semantics θ of the input S into the encoded information:

$$\mathbf{h}_i^{\text{topic}} = \tanh(\text{Linear}([\mathbf{h}_i, \theta])), \quad (6)$$

where $[\cdot]$ denotes the concatenation operator.

Causal Intervention for Predicting Counterfactuality. As mentioned in Section 3.3, we propose to debias hidden vectors from the imbalanced label bias. To this end, we set $p(d_{\text{CFD}}) = \frac{1}{|\mathcal{Y}|}$, where \mathcal{Y} is the set of groundtruth labels. Formally, the Eq. (4) becomes:

$$p(y|\text{do}(c)) = \frac{1}{|\mathcal{Y}|} \sum_{d_{\text{CFD}}} p(y|c, d_{\text{CFD}}). \quad (7)$$

Because of d_{CFD} , we need to incorporate the label information into $p(y|c, d_{\text{CFD}})$. We propose that information of each label exists in the hidden vectors of the inputs belonging to that label and denote such set of inputs for each label l as \mathcal{D}_l . Inspired by the prototypical network (Snell et al., 2017), we extract the information as follows:

$$\mathbf{h}_{[\text{CLS}]}^l = \frac{1}{|\mathcal{D}_l|} \sum_{S_j \in \mathcal{D}_l} \mathbf{h}_{j, [\text{CLS}]}. \quad (8)$$

Hereafter, we forward both the topic-oriented representation $\mathbf{h}_i^{\text{topic}}$ and the label information in $\mathbf{h}_{[\text{CLS}]}^l$ to the non-linear layer to classify the counterfactuality as:

$$\mathbf{h}'_{[\text{CLS}]} = \text{Linear} \left(\left[\mathbf{h}_{[\text{CLS}]}^{\text{topic}}, \text{Linear} \left[\{\mathbf{h}_{[\text{CLS}]}^l\}_{l \in \mathcal{Y}} \right] \right] \right), \quad (9)$$

$$p_{\text{CFD}} = p(y|\text{do}(c)) = \frac{1}{|\mathcal{Y}|} \sum_l \varphi(\mathbf{h}'_{[\text{CLS}]})^l, \quad (10)$$

where φ denotes the sigmoid function.

3.5 Training Strategy

Deconfounding NTM. To deconfound the NTM, we derive the Eq. (4) as:

$$p(\theta|\text{do}(\mathbf{x}_{\text{bow}})) = \sum_{d_{\text{TM}}} \frac{p(\theta, \mathbf{x}_{\text{bow}}|d_{\text{TM}}) \cdot p(d_{\text{TM}})}{p(\mathbf{x}_{\text{bow}}|d_{\text{TM}})}. \quad (11)$$

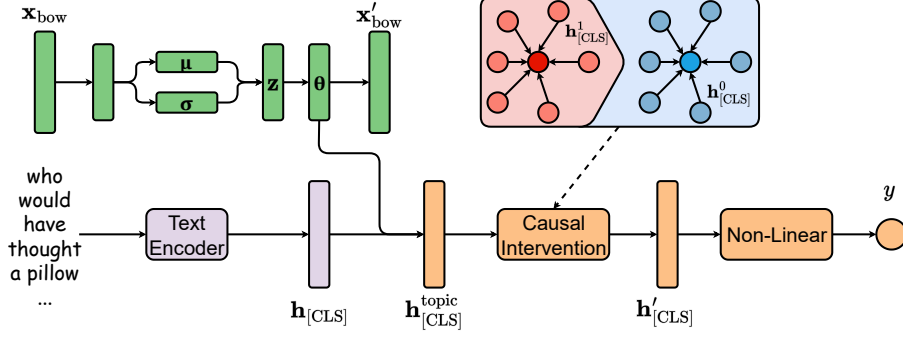


Figure 3: Illustration of the Topic-aware Causal Intervention Framework for Counterfactual Detection. Here the green component denotes the neural topic model, the purple component the text encoder, and the orange component our causal intervention operation for counterfactuality prediction.

In NTMs, topics are parameterized as word distributions (Blum and Haghtalab, 2016; Austin, 2011), similar to \mathbf{x}_{bow} . Hence, we conjecture that topic representation is a decomposed variant of each \mathbf{x}_{bow} , and we can only fully observe the distribution of the decompositions as in Figure 1 with the same number of times we retrieve \mathbf{x}_{bow} . Furthermore, as the training progresses, the output \mathbf{x}'_{bow} will converge to \mathbf{x}_{bow} . As such, we propose to approximate Eq. (11) following the propensity score modeling approach (Rosenbaum and Rubin, 1983):

$$p(\theta | \text{do}(\mathbf{x}_{\text{bow}})) \approx \sum_{d_{\text{TM}}} p(\theta, \mathbf{x}'_{\text{bow}}, d_{\text{TM}}) = \prod_i \frac{\phi_i \cdot \theta}{\|\phi_i\| \cdot \|\theta\|}, \quad (12)$$

where i refers to a word in \mathbf{x}_{bow} , and we empirically add the magnitude of θ . The denominator works as a normalizer to balance the magnitude of the variables.

Training Objective. Our framework jointly optimizes the Neural Topic Model and Counterfactual Detection (CFD) module. To train the CFD module, we employ the binary cross-entropy loss as:

$$\mathcal{L}_{\text{CFD}}(S, \mathbf{x}_{\text{bow}}, y) = -y \log p_{\text{CFD}} - (1 - y) \log(1 - p_{\text{CFD}}). \quad (13)$$

For the NTM, with the Eq. (12), we obtain the deconfounded evidence lower bound as:

$$\mathcal{L}_{\text{NTM}}(\mathbf{x}_{\text{bow}}) = \text{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} [\log p_{\phi}(\mathbf{x}_{\text{bow}}|\theta)] - \gamma \cdot \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[\sum_{i=1}^V \log \frac{\phi_i \cdot \theta}{\|\phi_i\| \cdot \|\theta\|} \right], \quad (14)$$

where the first term denotes the Kullback-Leibler divergence between the prior and posterior distribution, the second term the reconstruction error

of the output compared with the input, the third term the deconfounded objective in Eq. (12), V the vocabulary size, and γ the hyperparameter to control the deconfounding effect upon the training, respectively.

To conclude, our entire architecture is optimized with the linear combination of the loss functions \mathcal{L}_{NTM} and \mathcal{L}_{CFD} as:

$$\mathcal{L} = \mathcal{L}_{\text{CFD}} + \lambda_{\text{NTM}} \mathcal{L}_{\text{NTM}}, \quad (15)$$

where λ denotes the hyperparameter weight to scale the topic modeling component.

4 Experiments

4.1 Datasets

We evaluate on two prevalent datasets for the counterfactual detection task, SemEval-2020 (Yang et al., 2020) and Amazon-2021 (O’Neill et al., 2021). While SemEval-2020 comprises English documents, Amazon-2021 covers statements in three languages, English, Japanese, and German. For our experiments, we inherit the original train/val/test splits. To verify the generalizability of our methods, we measure our performance on two other bias-sensitive document analysis tasks, Paraphrase Identification with the MRPC dataset (Dolan and Brockett, 2005), and Implicit Sentiment Analysis (ISA) with CLIPeval from SemEval 2015 task 9 (Russo et al., 2015). These two tasks have been shown to sustain syntactic phrase and label biases (Li et al., 2020; Wang et al., 2022). The statistics of the datasets are provided in the Appendix. For evaluation metrics, we report Matthew’s Correlation Coefficient (MCC) (Boughorbel et al., 2017), the Accuracy (Acc), and F1 score.

4.2 Implementation Details

For the topic model, we select the topic number $T = 15$ based on the validation performance. Because at the beginning of the training process, the reconstructed output \mathbf{x}'_{bow} does not resemble the input \mathbf{x}_{bow} , we decide to adapt the linear warm-up strategy (Gilmer et al., 2021) with the number of warm-up steps $N_{\text{wp}} = 1000$ for the value of γ before fixing it to 0.25. We finetune two pretrained multilingual language models, mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2020) for the CFD task, and the monolingual BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) for the PI and ISA tasks. All variants are equipped with a linear layer on top of the pretrained language model. Our entire architecture is trained end-to-end on the A100 GPU, with the batch size of 16 and λ_{NTM} of 0.5 for 50 epochs, adopting Adam optimizer for the learning rate of 10^{-5} and L2 regularization equal 10^{-6} . For the counterfactual detection and paraphrase identification tasks, $\mathcal{Y} = \{0, 1\}$, meanwhile for the implicit sentiment analysis task, $\mathcal{Y} = \{-1, 0, 1\}$.

4.3 Baselines

As baselines, we compare our work against a wide variety of recent state-of-the-art bias-resolving causal intervention and data augmentation approaches for Counterfactual Detection: (i) **Stochastic Perturbation (SP)** (Wang et al., 2022), leveraging word perturbation to causally intervene the spurious effect of the language bias confounder; (ii) **Masking** (O’Neill et al., 2021), masking clue phrases in counterfactual detection to eliminate their effect upon the training; (iii) **Debiased Focal Loss (DFL)** (Karimi Mahabadi et al., 2020), de-emphasizing the loss contribution of easy biased examples and direct the model towards hard but less biased ones; (iv) **Product of Experts** (Karimi Mahabadi et al., 2020), aggregating the predictions of two models, one trained with the biased and the other with both biased and unbiased examples; (v) **Backtranslation** (O’Neill et al., 2021), a data augmentation method on the input level to increase the number of rare-label samples.

4.4 Comparison with State-of-the-arts

Results on Original Test Sets. We train and test the baselines and our model on the original test sets in Table 2. In the English variant of Amazon-2021 dataset, with mBERT we achieve an improved

accuracy of 1.5 points, and MCC score of 2.4 points with XLM-R. For German documents, our XLM-R outperforms the baseline using Backtranslation with 1.5 points, while our method adopted on mBERT enhances the SP approach with 1.0 point in MCC. On the Japanese version, where the language upholds syntactic and morphological features separate from English and German, our mBERT-based and XLM-R-based models accomplish absolute enhancements of 5.1 and 1.2 points in F1 metric, respectively, compared with DFL and SP, which are the best previous approaches.

On the SemEval-2020 dataset, which is at a larger scale and concerns diverse domains (O’Neill et al., 2021), our general performance is also auspicious. In particular, our mBERT system surpasses the DFL model by a mean MCC of 2.1 points. In addition, our XLM-R polishes the SP approach by 1.1 points of the F1 score. These results corroborate that our counterfactual detection model is able to cope with harmful confounding impacts of different biases, thus producing more generalizable representations to attain better performance.

Results on Balanced Test Sets. We randomly sample 500 samples from each class in the test set of SemEval-2020, then evaluate our method in Table 3. As can be seen, our method surpasses the best previous baseline, i.e. Backtranslation, with a significant margin of 1.5 points of accuracy for the mBERT variant, and surpasses DFL with 2.8 points of MCC for the XLM-R variant. These results verify that our causal intervention technique is able to mitigate the confounding effect of the class imbalance and makes the CFD model impartially consider the counterfactual and non-counterfactual choices when making prediction.

4.5 Zero-Shot Cross-lingual Evaluation

To clearly confirm whether our methods have the ability to deal with the bias of clue phrases, we conduct the zero-shot cross-lingual evaluation. In particular, we proceed to finetune the standard mBERT, XLM-R, and our counterfactual detection architectures on the Japanese portion of Amazon-2021 dataset, then directly validate the performance on the English portion, and similarly we finetune the models on the German training set and test them on the Japanese test set. We indicate the results in Table 4 and 5.

As can be observed, our model is capable of enhancing zero-shot cross-lingual counterfactual

Methods	Amazon-2021 (CD)									SemEval-2020 (CD)			MRPC (PI)			CLIPeVal (ISA)		
	En			De			Jp			En			En					
	Acc	MCC	F1	Acc	MCC	F1	Acc	MCC	F1	Acc	MCC	F1	Acc	MCC	F1	Acc	MCC	F1
mBERT/BERT	91.79	72.29	79.19	90.79	77.02	93.00	92.93	60.87	58.93	94.39	68.87	71.83	91.75	83.86	91.35	83.10	73.79	80.67
w/ DFL	93.88	81.70	81.30	91.11	79.58	93.47	<u>94.00</u>	<u>66.61</u>	<u>69.89</u>	<u>96.63</u>	<u>81.23</u>	<u>82.80</u>	92.15	84.62	91.69	<u>85.25</u>	<u>77.22</u>	<u>83.76</u>
w/ PoE	<u>94.03</u>	82.72	81.54	90.90	78.53	93.32	93.79	66.23	69.74	95.33	77.66	80.00	92.23	85.61	91.73	84.18	76.16	82.37
w/ Backtranslation	<u>94.03</u>	<u>83.07</u>	<u>81.89</u>	90.26	73.51	92.47	93.25	62.50	60.27	95.94	79.18	80.99	92.30	85.95	91.75	83.10	74.22	82.33
w/ Masking	93.43	78.60	81.01	91.43	79.79	93.89	93.68	64.28	68.02	95.81	78.73	80.84	—	—	—	—	—	—
w/ SP	93.63	81.19	81.21	<u>91.65</u>	<u>81.06</u>	<u>93.86</u>	93.61	62.76	66.83	95.08	77.35	79.02	<u>92.38</u>	<u>86.53</u>	<u>91.76</u>	84.44	76.97	83.13
Our Model	95.52	86.37	83.05	92.29	82.08	94.40	95.29	73.79	75.00	96.97	83.31	84.81	93.65	87.38	93.20	86.79	79.28	85.00
XLM-R/roBERTa	92.63	73.16	82.89	90.55	80.18	93.37	92.96	64.70	67.25	94.43	83.68	85.05	91.25	88.46	92.75	88.16	81.31	85.78
w/ DFL	94.66	85.22	83.46	90.84	80.32	93.53	94.40	74.21	75.68	96.41	84.43	85.56	93.92	88.99	94.05	<u>88.41</u>	81.64	85.89
w/ PoE	94.52	84.90	83.09	90.86	80.58	93.74	94.12	73.23	75.59	96.43	84.55	85.87	93.50	88.91	93.75	87.87	80.80	84.99
w/ Backtranslation	94.95	85.85	83.81	<u>91.79</u>	<u>81.02</u>	<u>93.99</u>	94.33	73.75	75.67	95.86	84.36	85.43	93.34	88.78	93.62	87.60	80.48	84.86
w/ Masking	94.20	83.75	82.96	90.21	79.75	93.25	94.57	74.44	75.73	95.79	84.22	85.11	—	—	—	—	—	—
w/ SP	<u>95.40</u>	<u>86.35</u>	<u>83.87</u>	91.18	80.76	93.87	<u>94.87</u>	<u>75.03</u>	<u>76.85</u>	<u>96.77</u>	<u>85.18</u>	<u>86.34</u>	<u>94.60</u>	<u>89.22</u>	<u>94.17</u>	<u>88.41</u>	<u>81.69</u>	<u>87.03</u>
Our Model	96.85	88.74	84.44	92.51	82.49	94.57	95.82	76.01	77.97	97.44	86.09	87.44	95.55	91.05	95.14	89.49	83.55	88.07

Table 2: Numerical results on original test sets of the Counterfactual Detection (CD), Paraphrase Identification (PI), and Implicit Sentiment Analysis (ISA) tasks. We respectively bold and underline the best and second-to-best results.

Methods	SemEval-2020		
	Acc	MCC	F1
mBERT	85.75	74.21	83.55
w/ DFL	91.48	83.53	90.94
w/ PoE	90.45	82.01	89.60
w/ Backtranslation	91.68	83.21	91.54
w/ Masking	89.53	80.28	88.52
w/ SP	90.60	81.92	89.94
Our model	93.13	86.52	92.84
XLM-R	88.95	79.69	88.04
w/ DFL	92.78	86.01	92.38
w/ PoE	91.33	83.50	90.66
w/ Backtranslation	92.48	85.94	92.49
w/ Masking	90.60	82.15	89.83
w/ SP	89.40	80.13	88.25
Our model	94.33	88.83	94.14

Table 3: Numerical results on balanced test sets of the CFD task on the SemEval-2020 dataset.

Models	Jp → En			De → Jp		
	Acc	MCC	F1	Acc	MCC	F1
mBERT	91.83	49.78	52.81	80.84	40.54	44.92
Our Model	93.40	59.41	60.71	91.54	50.10	54.34

Table 4: Cross-lingual Zero-shot mBERT results on the Amazon-2021 dataset.

detection capacity of both mBERT and XLM-R, surpassing mBERT with a large margin of 1.6 and XLM-R with 1.1 points in accuracy for the English test set. For the Japanese test set, we outperform mBERT and XLM-R with 9.6 points of F1, and 9.7 points of MCC, respectively. These results substantiate that our methods can mitigate the clue phrase bias in the language models.

4.6 Adaptability to Other Bias-sensitive Tasks

Experimental results in PI and ISA tasks are given in Table 2. For the MRPC dataset, our BERT model performance exceeds the one of the SP method by 1.3 points in accuracy, and 1.4 points in F1. With respect to the RoBERTa backbone, we also surpass

Models	Jp → En			De → Jp		
	Acc	MCC	F1	Acc	MCC	F1
XLM-R	92.70	61.55	62.83	87.87	45.82	50.49
Our Model	93.85	62.79	64.35	89.19	55.56	51.67

Table 5: Cross-lingual Zero-shot XLM-R results on the Amazon-2021 dataset.

Methods	Acc	MCC	F1
Our Model	95.52	86.37	83.05
w/o Debaised CFD Objective	94.63	83.75	82.96
w/o Deconfounded Topic Model Objective	94.33	83.29	82.52
w/o Neural Topic Model	93.43	80.59	82.40

Table 6: Results from ablating different deconfounding components on the English Amazon-2021 dataset.

the SP method by 1.8 points of MCC, and 1.0 point of F1. Regarding the CLIPeVal dataset, integrating our approaches into BERT and RoBERTa extends the performance with 2.1 points in MCC, and 1.1 points in accuracy, respectively.

Those aforementioned results have shown that our methods have the capability of tackling biases in not only counterfactual detection but also other natural language understanding tasks.

4.7 Ablation Study

Effect of Deconfounding Components. In this ablation, we experiment with removing each component that helps the model deconfound. Particularly, we train and test the ablated mBERT on the English portion of the Amazon-2021 dataset. As shown in Table 6, solely employing one of the elements does enhance the overall counterfactual recognition, but being less effective than the joint approach. Without combining the deconfounding mechanisms, the model might not be able to cope with multiple biases.

Effect of Global Semantics. Here, we investigate the performance of our method when utiliz-

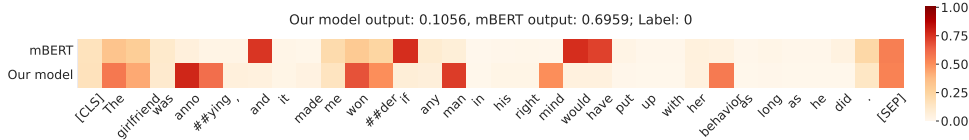


Figure 4: Attention weights of the [CLS] token to all other words, and output scores of mBERT and Our Model. The score is in range [0, 1]. The input: “The girlfriend was annoying, and it made me wonder if any man in his right mind would have put up with her behavior as long as he did.”

Methods	Acc	MCC	F1
XLM-R + NTM	92.51	82.49	94.57
XLM-R + PFA	91.86	81.09	94.09
XLM-R + LDA	91.97	81.83	94.25

Table 7: Ablation results with various types of global semantics on the German Amazon-2021 dataset.

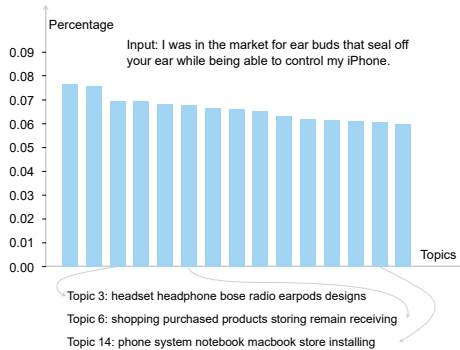


Figure 5: Topic Percentages and inferred Top Topics from Figure 1 after Causal Intervention.

ing conventional topic models. We consider two choices, i.e. Poisson Factor Analysis (PFA) and Latent Dirichlet Allocation (LDA), and finetune the XML-R model on the German subset in Amazon-2021. As can be seen in Table 7, NTM burnishes the counterfactual detector more effectively than traditional topic models. More results on the two ablation experiments can be found in the Appendix.

4.8 Case Study

Impact of Causal Intervention on Attention Logits. Here we randomly select one example from Table 1 and visualize the average attention scores in the heads of all layers of the [CLS] token to the remaining words. As shown in Figure 4, whereas mBERT’s [CLS] strongly pays attention to clue phrases “if” and “would have”, our model distributes the attention impartially and emphasizes content words, such as “annoying”, “man”, and “behavior”. This could help to explain our more reasonable prediction than mBERT. We provide attention visualizations of other examples in the Appendix. These visualizations demonstrate that our approach can resolve the confounding influence of clue phrases and improve model prediction.

Impact of Causal Intervention on Topic Distri-

bution. In Figure 5, we obtain topic representation from our neural topic model for the document of the Amazon-2021 dataset in Figure 1, and then count the percentage of documents sharing the top topic, i.e. possessing the largest likelihood. Different from Figure 1, our deconfounded topic model does not lean towards a subset of topics to assign top probabilities. Moreover, all three leading topics reveal the semantics of the document, which concerns *headset*, *shopping*, and *phone*. These results demonstrate that our approach is capable of resolving the topic bias phenomenon to produce faithful global semantics for counterfactual detection.

5 Conclusion

In this paper, we propose a causal intervention framework that discovers biases in the counterfactual detection problem. In order to cope with clue phrase, topic, and label biases, we propose to utilize global semantics and extend the training strategy with deconfounding training objectives. Comprehensive experiments demonstrate that our model can ameliorate detrimental influences of biases to polish previous state-of-the-art baselines for not only the counterfactual detection but also other bias-sensitive NLU tasks.

6 Limitations

We consider the following two limitations as our future work: (1) Extend the problem to circumstances with multiple observable confounding variables. The problem will become more complex if additional confounding factors are explicitly taken into account. Studying such complex scenario is potential to enhance the applicability and our understanding towards the proposed debiasing technique; (2) Explore the impact of causal intervention on generative tasks. We have only verified the effectiveness of causal intervention in discriminative language models. Whether the effectiveness applies for generative tasks such as machine translation, document summarization, etc., remains an open problem and interesting research direction.

References

- Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.
- Avrim Blum and Nika Haghtalab. 2016. Generalized topic modeling. *arXiv preprint arXiv:1611.01259*.
- Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLoS one*, 12(6):e0177678.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.
- Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Dahl, Zachary Nado, and Orhan Firat. 2021. A loss curvature perspective on training instability in deep learning. *arXiv preprint arXiv:2110.04369*.
- Rabeeh Karimi Mahabadi, James Henderson, et al. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, CONF*.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, pages 2410–2419. PMLR.
- Cong-Duy Nguyen, Thong Nguyen, Duc Vu, and Anh Tuan Luu. 2023a. Improving multimodal sentiment analysis: Supervised angular margin-based contrastive learning for enhanced fusion representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14714–14724.
- Cong-Duy Nguyen, Thong Nguyen, Xiaobao Wu, and Anh Tuan Luu. 2024a. Kdmcse: Knowledge distillation multimodal sentence embeddings with adaptive angular margin contrastive learning. *arXiv preprint arXiv:2403.17486*.
- Cong-Duy Nguyen, The-Anh Vu-Le, Thong Nguyen, Tho Quan, and Anh-Tuan Luu. 2023b. Expand bert representation with visual information via grounded language learning with multimodal partial alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5665–5673.
- Thong Nguyen, Yi Bin, Xiaobao Wu, Xinshuai Dong, Zhiyuan Hu, Khoi Le, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024b. Meta-optimized angular margin contrastive framework for video-language representation learning. *arXiv preprint arXiv:2407.03788*.
- Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024c. Video-language understanding: A survey from model architecture, model training, and data perspectives. *arXiv preprint arXiv:2406.05615*.
- Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *Advances in neural information processing systems*, 34:11974–11986.
- Thong Nguyen, Anh Tuan Luu, Truc Lu, and Tho Quan. 2021. Enriching and controlling global semantics for text summarization. *arXiv preprint arXiv:2109.10616*.
- Thong Nguyen, Xiaobao Wu, Xinshuai Dong, Khoi M Le, Zhiyuan Hu, Cong-Duy Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024d. Read-pvla: Recurrent adapter with partial video-language alignment for parameter-efficient transfer learning in low-resource video-language modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18824–18832.
- Thong Nguyen, Xiaobao Wu, Xinshuai Dong, Anh Tuan Luu, Cong-Duy Nguyen, Zhen Hai, and Lidong Bing. 2023c. Gradient-boosted decision tree for listwise context model in multimodal review helpfulness prediction. *arXiv preprint arXiv:2305.12678*.
- Thong Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy Nguyen, See Kiong Ng, and Anh Luu. 2023d. Demaformer: Damped exponential moving average transformer with energy-based modeling for temporal language grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3635–3649.

- Thong Nguyen, Xiaobao Wu, Anh-Tuan Luu, Cong-Duy Nguyen, Zhen Hai, and Lidong Bing. 2022. Adaptive contrastive learning on multimodal transformer for review helpfulness predictions. *arXiv preprint arXiv:2211.03524*.
- Thong Thanh Nguyen, Zhiyuan Hu, Xiaobao Wu, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024e. Encoding and controlling global semantics for long-form video question answering. *arXiv preprint arXiv:2405.19723*.
- Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11103–11111.
- Thong Thanh Nguyen, Xiaobao Wu, Xinshuai Dong, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024f. Topic modeling as multi-objective optimization with setwise contrastive learning. In *The Twelfth International Conference on Learning Representations*.
- James O’Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish i would have loved this one, but i didn’t—a multilingual dataset for counterfactual detection in product review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7092–7108.
- Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.
- Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 443–450.
- Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.
- Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 654–658.
- Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524.
- Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Causal intervention improves implicit sentiment analysis. *arXiv preprint arXiv:2208.09329*.
- Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031.
- Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liang-Ming Pan, and Anh Tuan Luu. 2023a. Infotm: A mutual information maximization perspective of cross-lingual topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13763–13771.
- Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023b. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, pages 37335–37357. PMLR.
- Xiaobao Wu, Xinshuai Dong, Liangming Pan, Thong Nguyen, and Anh Tuan Luu. 2024a. Modeling dynamic topics in chain-free fashion by evolution-tracking contrastive learning and unassociated word exclusion. *arXiv preprint arXiv:2405.17957*.
- Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024b. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18.
- Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024c. Fastopic: A fast, adaptive, stable, and transferable topic modeling paradigm. *arXiv preprint arXiv:2405.17978*.
- Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024d. On the affinity, rationality, and diversity of hierarchical topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19261–19269.
- Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. 2020. Semeval-2020 task 5: Counterfactual recognition. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 322–335.
- Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. 2022. Certified robustness against natural language attacks by causal intervention. *arXiv preprint arXiv:2205.12331*.

A Attention Visualization

In this section, we visualize the attention weights of the [CLS] token to the words of the examples in Table 1.

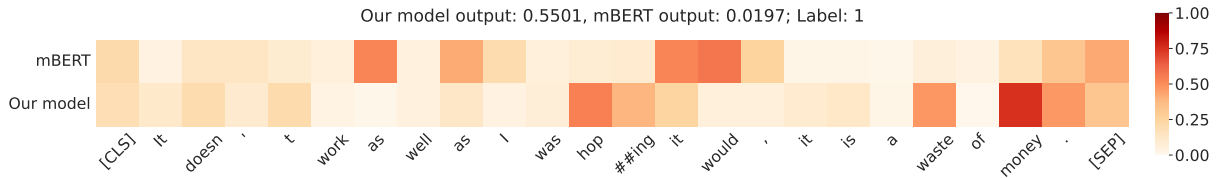


Figure 6: Attention weights of the [CLS] token to all other words, and output scores of mBERT and our model. The score is in range $[0, 1]$. The input: “*It doesn’t work as well as I was hoping it would, it is a waste of money.*”

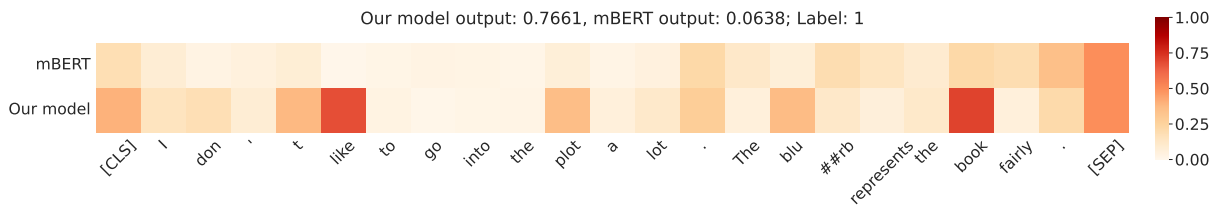


Figure 7: Attention weights of the [CLS] token to all other words, and output scores of mBERT and our model. The score is in range $[0, 1]$. The input: “*I don’t like to go into the plot a lot. The blurb represents the book fairly.*”

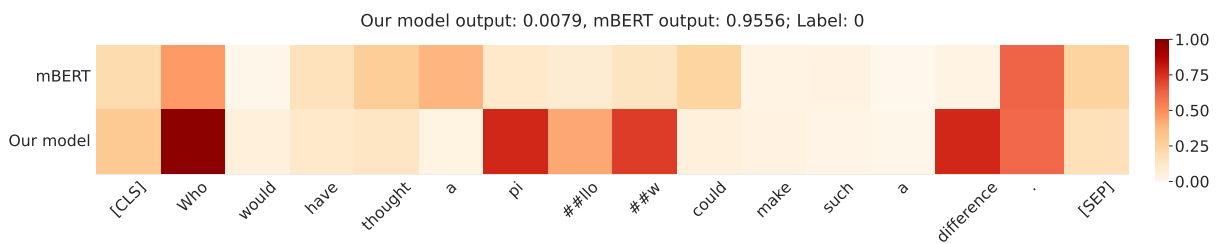


Figure 8: Attention weights of the [CLS] token to all other words, and output scores of mBERT and our model. The score is in range $[0, 1]$. The input: “*Who would have thought a pillow could make such a difference.*”

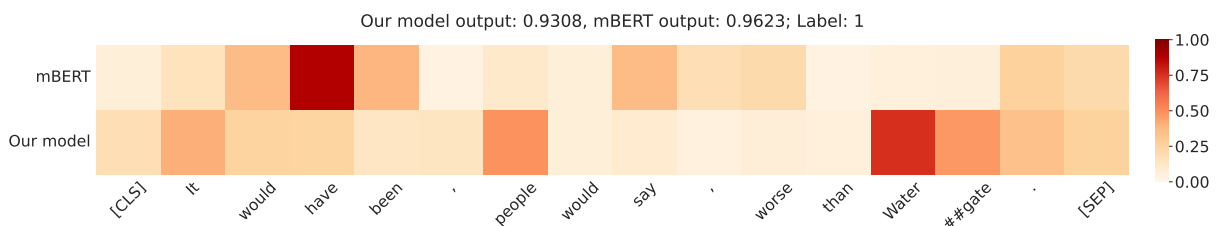


Figure 9: Attention weights of the [CLS] token to all other words, and output scores of mBERT and our model. The score is in range $[0, 1]$. The input: “*It would have been, people would say, worse than Watergate.*”

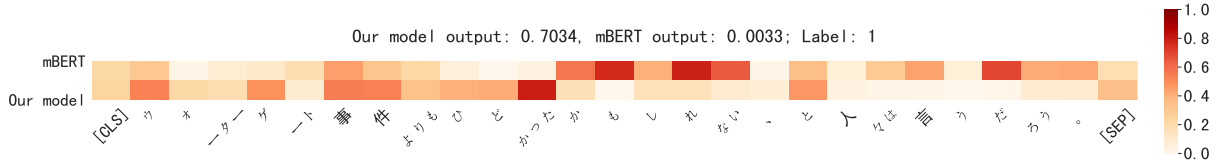


Figure 10: Attention weights of the [CLS] token to all other words, and output scores of mBERT and our model. The score is in range [0, 1]. The input: “ウォーターゲート事件よりもひどかったかもしれない、と人々は言うだろう。”

B Dataset Statistics

In this section, we present the statistics of all the datasets pertaining to Counterfactual Detection, Paraphrase Identification, and Implicit Sentiment Analysis tasks.

Dataset	Variant	Train	Val	Test	#Pos	#Neg	#Neutral	#Pos in Test	#Neg in Test	Total
Amazon-2021	En	4018	335	670	954	4069	-	131	539	5023
	De	5600	466	934	4840	2160	-	650	284	7000
	Jp	5600	466	934	667	6333	-	96	838	7000
SemEval-2020	En	13000	-	7000	2192	17808	-	738	6262	20000
MRPC	En	49184	2000	2000	23493	29691	-	907	1093	53184
CLIPeval	En	1347	-	371	580	796	342	216	155	1718

Table 8: Statistics of the Datasets.

C Additional Ablation Studies

Impact of Deconfounding Components. We compare our model with its ablated variants in all subsets of the Amazon-2021 dataset. As can be observed in Table 9, jointly utilizing deconfounded neural topic model and debiased objective can tackle the clue phrase, label, and topic biases, leading to the largest overall improvement.

Methods	En			De			Jp		
	Acc	MCC	F1	Acc	MCC	F1	Acc	MCC	F1
Our Model	95.52	86.37	83.05	92.29	82.08	94.40	95.29	73.79	75.00
w/o Debiased CFD objective	94.63	83.75	82.96	92.15	81.49	94.36	94.87	73.13	73.61
w/o Deconfounded Topic Model objective	94.33	83.29	82.52	91.94	81.31	94.19	94.80	72.78	71.66
w/o Neural Topic Model	93.43	80.59	82.40	91.76	81.20	94.05	94.72	72.60	71.09

Table 9: Results of subsequently pruning deconfounding components on the Amazon-2021 dataset.

Impact of Global Semantics. In addition to the results in Table 7, we execute our model with different topic models on other languages of the Amazon-2021 dataset. The results are shown in Table 10.

Methods	En			De			Jp		
	Acc	MCC	F1	Acc	MCC	F1	Acc	MCC	F1
XLM-R + NTM	96.85	88.74	84.44	92.51	82.49	94.57	95.82	76.01	77.97
XLM-R + PFA	96.63	87.19	83.97	91.86	81.09	94.09	95.72	75.46	77.68
XLM-R + LDA	96.40	87.11	83.10	91.97	81.83	94.09	95.18	75.07	77.53

Table 10: Results with heterogeneous topic models on the Amazon-2021 dataset.

UD for German Poetry

Stefanie Dipper and Ronja Laarmann-Quante

Ruhr University Bochum, Germany

Faculty of Philology

Department of Linguistics

{firstname.lastname}@ruhr-uni-bochum.de

Abstract

This article deals with the syntactic analysis of German-language poetry from different centuries. We use Universal Dependencies (UD) as our syntactic framework. We discuss particular challenges of the poems in terms of tokenization, sentence boundary recognition and special syntactic constructions. Our annotated pilot corpus currently consists of 20 poems with a total of 2,162 tokens, which originate from the PoeTree.de corpus. We present some statistics on our annotations and also evaluate the automatic UD annotation from PoeTree.de using our annotations.

1 Introduction

For many years, poetry played a rather subordinate role in the field of Natural Language Processing. With the emergence of the field of Digital Humanities, the interest in computational methods for philological questions has increased considerably.

It is often assumed that the particular repetitive structures and patterns of poems serve, among other things, to support memorization. These structures prominently include meter and the rhyme scheme. Many computational approaches to the analysis of poetry have focused on these features, such as the work by Bobenhausen and Hammerich (2015); Haider (2023, 2021); Delente and Renault (2017), who automatically analyze phonological features like meter and rhyme in English, German and French poetry (also see the overview in De Sisto et al., 2024).

On a more abstract level, syntactic patterns can also support memorization, e.g. by using the same syntactic structure several times. Lee and Kong (2012) investigate and compare such constructions in poems of two Chinese poets under the name ‘parallel couplets’, which refers to semantic or syntactic correspondences between two lines.

Another syntactic phenomenon that is specific to poems is unusual word order, which often results from metrical constraints. A special case is the enjambment, in which the elements of a syntactic phrase are separated by spreading them over two lines, whereby special retarding effects can be achieved. Enjambment has been examined by Ruiz Fabo et al. (2017) and Hussein et al. (2018).

Syntactic annotations of poems are also interesting for other reasons. For instance, DeHass (2024) uses them to compare paratactic vs. hypotactic style in Latin colloquial texts, poetry, and prose. Syntax annotations can in general facilitate access to the content of a text and, e.g., allow for easy retrieval of (syntactically and semantically) related units such as entities and their properties, or events and their participants, as suggested by Bamman (2020) for literary texts.

For a long time, such studies could only be carried out on a rather small amount of data that has been manually annotated by experts. Examples of such manually-created poetry treebanks are the York-Helsinki parsed corpus of Old English poetry (YCOEP)¹ (Pintzuk and Leendert, 2001), the Ancient Greek and Latin Dependency Treebank (AGLDT)² (Bamman and Crane, 2011), which comprises a subset of poetic texts, and the treebank of Classical Chinese poems (Lee and Kong, 2012).

Only recently has there been work on the *automatic* syntax analysis of poems and the first poetry treebanks have been generated automatically, such as the diachronic treebank of Spanish Sonnets (Ruiz Fabo et al., 2017), the treebank of Classical Arabic poetry (Al-Ghamdi et al., 2021), or the PoeTree treebank with poems in ten different languages (Plecháč et al., 2024). All three treebanks were automatically annotated by generic parsers

¹<https://www-users.york.ac.uk/~lang18/pcorpus.html>

²http://perseusdl.github.io/treebank_data/

that were not specifically trained for poems because no training data was available (see Section 4.2 for evaluation results).

Our work aims to contribute to the creation of training and evaluation data for the syntactic analysis of poetry. Our paper deals with the analysis of German-language poems from different centuries. We use the framework of Universal Dependencies (UD, de Marneffe et al., 2021) and discuss some difficulties for annotation that arise due to the peculiarities of poems. We manually annotate a pilot corpus of 20 poems with 2,162 tokens, following the guidelines proposed by Dipper et al. (2024), who define a set of customized labels for German.

Our data comes from the PoeTree corpus, that has been automatically annotated with UD trees by UDPipe (Straka, 2018). We evaluate the UD trees by means of our manually created annotations. For the evaluation, we map the customized labels back to UD labels. The main contributions of this paper are:

- A pilot UD treebank of currently 20 German-language poems from different centuries, which are available under a free license.³
- An in-depth discussion of the specific challenges of (German) poetry.
- A first evaluation of the automatic analyses from UDPipe.

2 Data

The data for our study comes from the PoeTree corpus, which consists of more than 330,000 poems with 89,000,000 tokens from 10 European languages (Plecháč et al., 2023; Plecháč et al., 2024).⁴ All poems have been annotated automatically with UD-style dependencies using UDPipe 2.0 (Straka, 2018). However, only the annotations of the Czech-language subcorpus have already been evaluated (Cinková et al., 2024) (see Section 4.2).

The German-language sub-corpus of the PoeTree corpus, called PoeTree.de (Bobenhausen and Hammerich, 2015; Haider, 2021) consists of 74,000 poems. The automatic annotations comprise lemma, inflection features, universal part-of-speech (POS) tags, language-specific POS accord-

³The corpus is available at <https://gitlab.ruhr-uni-bochum.de/vamos-cl/ud-for-german-poetry>.

⁴Version 0.0.2, <https://versologie.cz/poetree/>.

	Mean \pm SD	Total
Tokens	108.1 \pm 92.8	2,162
Lines	15.7 \pm 15.1	314
Stanzas	4.9 \pm 5.7	97

Table 1: Statistics on the annotated German sub-corpus: mean number and standard deviation of tokens, lines and stanzas per poem.

ing to the STTS tagset (Schiller et al., 1999) and UD-style dependency relations.

We randomly selected 20 poems from PoeTree.de, see Table 1 for an overview and Table 6 in Appendix A for more details. Standard deviations are very large and show that the poems differ greatly with regard to their size. For annotating dependency relations, we use the manual annotation tool INCEPTION (Klie et al., 2018). Each poem was annotated once, by one of the authors, and difficult cases were discussed together.

3 Poetry-Specific Issues

In this section, we discuss selected special features of poetry that pose a challenge for automatic language processing. The focus is on German-language poetry; many of the challenges, however, also arise for poetry in other languages.

3.1 Word forms and tokenization

Poetry belongs to non-standard language data in several respects. For example, capitalization is often handled differently than in standard language, e.g., some poems are written entirely in lower case. In the poems of our corpus, the first word of a line is usually capitalized, whereas in standardized spelling in German only sentence beginnings and nouns are capitalized, so that the unusual capitalization sometimes results in incorrect POS tags. For instance, in (1) the verb *schreibe* ‘write’ occurs twice. The first (lowercase) instance is correctly tagged in the corpus with VVFIN (finite verb), the second (capitalized) instance is incorrectly tagged with NN (noun).

- (1) *Da sitz ich am Tisch und schreibe,
Schreibe wie mir es glückt,
‘I sit at the table and write, write the way I
manage.’* Source: 00-1734-0000-0002-9F08-4#0⁵

⁵ID as provided in the PoeTree.de corpus.

Due to the meter, word forms can be shortened (elision) or lengthened (epenthesis). Elisions are often marked by an apostrophe, which can be problematic for tokenization. For example, the elision in *heil'gen* ‘holy’ in (2), which stands for *heiligen*, has been split into three tokens *heil*, apostrophe and *gen* in PoeTree.de.⁶ We annotate such partial tokens with the UD relation *goeswith*.

- (2) *Liebste, laß in Dir die Schauer*
Weben dieser heil'gen Nacht,
 ‘Dearest, let the shivers of this holy night
 weave in you,’ Source: dta.poem.21583

Another difficulty is the UD treatment of contracted prepositions plus articles, which are very common in German, e.g. *am* \approx *an dem* ‘at the’. In UD style (and in the PoeTree data), such contractions are split and treated as two words, each annotated with its own UD relation (cf. Grünewald and Friedrich, 2020).⁷ In (manual) annotation of poetry, this procedure is problematic: on the one hand, the meter is no longer correct due to the inserted syllable, and on the other hand, there are differences in meaning between the contracted and the split form, i.e., one form cannot be replaced equivalently by the other (see, e.g., Cieschinger, 2016). Moreover, split forms are highly marked in German and make up less than 10% of the occurrences in a newspaper and a web corpus (Cieschinger, 2016, p. 6), i.e., the UD tokenization applied in PoeTree.de makes it harder to understand the poems. Annotators often have to re-merge the preposition and the article mentally when annotating in order to understand the meaning of a poem correctly. For example, the phrase *aufs Neue* ‘anew’ – shortened to *aufs neu*’ in the poem line shown in (3) – is rendered as *auf das neu*’ ‘on the new’ in PoeTree.de. The line with the split form could mean something like ‘always hope for the new’, which is clearly not the correct meaning and does not fit the context.

- (3) *Und was dir fehlschlug, hoffe stets aufs neu*
 ‘And what you have failed, always hope anew’
 Source: 00-1734-0000-0002-B719-B#0

⁶The partial tokens exist as independent words: *heil* means ‘sound’ and *gen* is an obsolete preposition meaning ‘towards’. Presumably this leads to the incorrect tokenization.

⁷Also see the distinction between tokenization and word segmentation in the UD Guidelines <https://universaldependencies.org/u/overview/tokenization.html>: Tokens are defined orthographically in UD and are marked by whitespace. Words are defined syntactically and, consequently, clitics are split off and contractions are undone.

Note that the original form *aufs* ‘on the’ is recorded in the ‘multiword’ entry in the JSON format of the PoeTree.de treebank and could be used in the dependency annotation. However, a new, complex dependency label would then have to be introduced for the contracted forms (e.g. case+det). This would deviate from the UD guidelines, though, and the automatically generated analyses of PoeTree could no longer be evaluated directly, which is why we have annotated the split forms.⁸

3.2 Sentence boundaries

Another feature is the non-standard punctuation of poetry, which, together with the unusual capitalization, means that sentence boundaries are often not correctly determined and, for example, arguments or modifiers are not located in the same sentence segment⁹ as their head. This causes problems for syntax analysis, since dependencies are usually only marked within a sentence. In our manual annotations we have treated incorrect sentence segments as follows:

1. If there are several sentences in one sentence segment, link them by the relation *parataxis*.
2. If two incomplete fragments are distributed over two sentence segments, each fragment is annotated as usually, but only those relations that apply within the segment are annotated. Note that such fragments are not linked to other parts of the sentence with the relation *parataxis*, so that there can be several roots within a sentence segment in our corpus.

An example is given in (4), which consists of one sentence. The double slashes “//” indicate the (incorrect) sentence boundaries as defined in PoeTree.de. This means that several central relations are missing, see the annotations shown in Figure 1: *Meinem Herzen* ‘to my heart’ is governed by the verb *reiche* ‘give’ as a dative object, *von der Schlange* ‘from the snake’ belongs to the verb *gebissen* ‘bitten’.

⁸In some UD treebanks, e.g. in the German-LIT treebank (Salomoni, 2017, 2021) or in the NArabizi treebank (Seddah et al., 2020), such contracted forms are preserved in their original form and composite POS tags are used, e.g. P+DET in NArabizi. However, at the dependency level, these forms are annotated by the label *case* only, so that the information on the determiner is not represented in the dependency relations.

⁹We refer to the (potentially incorrect) sentences as defined in the PoeTree.de corpus as “sentence segments”.

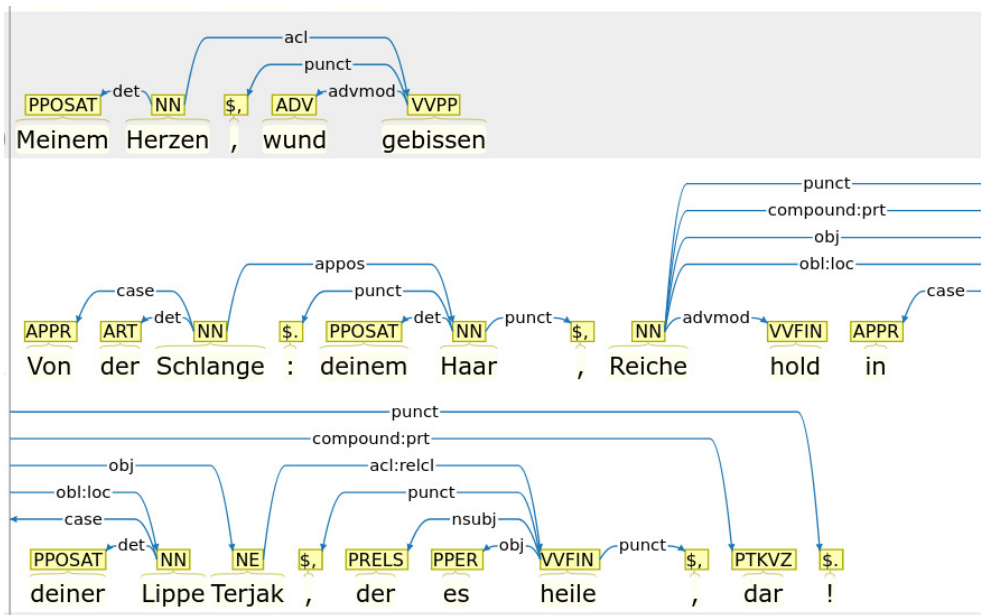


Figure 1: Annotation of the fragments of Ex. (4); screenshot of INCEpTION.

- (4) *Meinem Herzen, wund gebissen //
Von der Schlange: deinem Haar,
Reiche hold in deiner Lippe
Terjak, der es heile, dar!*
‘To my heart, bitten sore from the snake, [i.e.]
your hair, give in your lip Terjak [a medical
concoction], which heals it!’
Source: 00-1734-0000-0003-2E39-A#0

In Example (4) the fragments each form a complete subtree, with the roots *gebissen* ‘bitten’ and *Schlange* ‘snake’. In other cases, however, the heads of the fragments can also be in the other sentence segment, so that individual words then remain unrelated. For example, in (5) the verb head *tragen* ‘bear’ only occurs in the second sentence segment, so that its dependents in the first segment – the comma, the conjunction *dass* ‘that’ and the prepositional phrase *auf der Wiese* ‘in the meadow’ – remain unrelated, see the annotations in Figure 2.

- (5) *Erhebe dich, dass auf der Wiese //
Durch deines Wuchses hohes Streben
Zipressenbäume Früchte tragen*
‘Arise, so that in the meadow through your
growth’s high aspiration cypress trees bear
fruit’ Source: 00-1734-0000-0003-28BA-1#0

A final example is (6), in which the repeated instances of *Geduld!* ‘patience’ are separated by a (false) sentence boundary due to the exclamation mark.

- (6) *Geduld! //
Geduld! – die ew’gen Sterne gehn
Doch ihren Pfad.*
‘Patience! Patience! – the eternal stars go but
their path.’ Source: 00-1734-0000-0002-B719-B#0

3.3 Poetry-specific constructions

Repetitions Repetitions are a typical feature of poetry. We have already seen examples in (1) and (6), another example is (7), where *mein Herz* ‘my heart’ is repeated three times in a row. We annotate such repetitions like lists, but with a special label, *list:rep*.¹⁰

- (7) *Das Meer hat seine Perlen,
Der Himmel hat seine Sterne,
Aber mein Herz, mein Herz,
Mein Herz hat seine Liebe.*
‘The sea has its pearls, the sky has its stars,
but my heart, my heart, my heart has its love’
Source: dta.poem.10555

Anacoluthon Sentence interruptions (*anacolutha*) are also a typical feature. In Example (8), the first and third lines each begin with a free relative clause (*Die noch schlafen* ‘who still sleep’, *Die noch keimen* ‘who still germinate’), which is not integrated into the surrounding clauses and whose referent therefore remains vague.

¹⁰This label is a customized UD label using the notation *universal:customized* (see for example the already existing *aux:pass* for passive auxiliaries, de Marneffe et al., 2021).

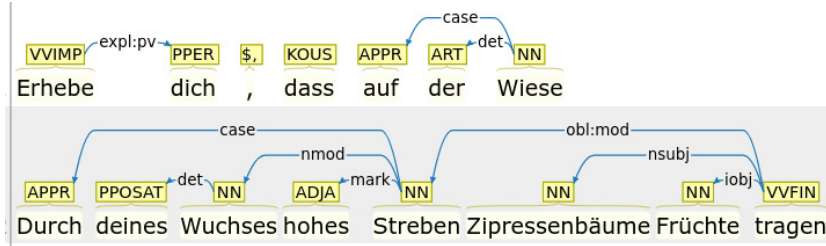


Figure 2: Partial annotation of the fragments of Ex. (5); screenshot of INCEpTION.

- (8) *Die noch schlafen, aus den Wäldern
Rauscht's wie leiser Vogelsang,
Die noch keimen, von den Feldern
Blüht's wie Duft das Thal entlang.*
‘Who still sleep, from the woods it rustles like
soft birdsong, who still germinate, from the
fields it blossoms like fragrance along the
valley.’ Source: dta.poem.21583

We annotate such interruptions with the label *parataxis*, as proposed by Paccosi et al. (2023) (see their example (3)). However, *parataxis* captures a rather heterogeneous class of relations, such as unconnected sentences and also parentheses. We currently also annotate cases in which a sentence boundary is missing (see Section 3.2) with this label. Additional subtypes for differentiation should possibly be defined here.

4 Results

In this section, we present selected statistics from our annotations. In the second part, we evaluate the automatic annotations of PoeTree.de.

4.1 Selected statistics

We first compare the distributions of the different labels with the two corpora for modern German from Dipper et al. (2024), which were annotated according to the same UD scheme. These corpora originate from the GSD treebank and contain modern news and reviews. Since all three corpora are rather small,¹¹ we truncate all extended labels (e.g. *obl:arg*) and map them to the basic labels (*obl*).

Table 2 shows the 10 most frequent labels in the three text types. Six of the labels (*punct*, *det*, *nsubj*, *advmod*, *case*, *obl*) occur in all three text types and their distributions are rather similar. Coordinations (labels *conj* and *cc*) are typical for reviews and poetry, while they make up less than 3% of the labels in news. The label *flat*, which

¹¹News: 100 sentences with 1,872 tokens; reviews: 100 sentences with 1,341 tokens.

	News	Reviews	Poetry
<i>punct</i>	16.2	<i>punct</i> 13.5	<i>punct</i> 16.9
<i>det</i>	13.5	<i>advmod</i> 12.5	<i>det</i> 14.4
<i>nsubj</i>	8.6	<i>det</i> 11.0	<i>nsubj</i> 10.2
<i>case</i>	8.6	<i>nsubj</i> 9.3	<i>advmod</i> 8.6
<i>advmod</i>	7.1	<i>case</i> 8.1	<i>case</i> 6.9
<i>obl</i>	5.9	<i>obl</i> 6.0	<i>obl</i> 6.3
<i>nmod</i>	4.8	<i>obj</i> 4.6	<i>obj</i> 5.2
<i>amod</i>	4.7	<i>conj</i> 4.4	<i>conj</i> 4.4
<i>flat</i>	3.8	<i>cc</i> 4.3	<i>cc</i> 3.9
<i>aux</i>	3.7	<i>aux</i> 4.0	<i>amod</i> 3.2
Total	1,772	1,241	1,962

Table 2: The 10 most frequent dependency labels in the poetry, news and reviews data, along with their relative frequencies (ignoring root and unspecified labels). The last line specifies the total number of labels in the respective data set.

is mainly used to annotate complex proper names, is typical for news; in reviews it occurs with 1.1%, in poetry not at all. The label *aux*, which is used to annotate modal verbs and auxiliaries, occurs in poetry with 2.3%, i.e. less often than in reviews and news.

We also compare the distributions of all labels using Spearman’s rank correlation coefficient r , which compares the ranks of the labels in the three text types, and the Jensen-Shannon distance, which is a symmetric and normalized variant of the Kullback-Leibler divergence and compares the probabilities of the labels with each other.¹²

Table 3 shows that with both measures, the text types news and reviews are very similar to each other (high correlation, small distance), while news and poetry are clearly less similar to each other. The reviews occupy a position between the other two text types. The data sets can therefore be arranged as follows based on their similarities:

¹²Spearman’s rank correlation coefficient and Jensen-Shannon distance were calculated with SciPy, <https://docs.scipy.org>.

	News	Reviews	Poetry
News	1		
Reviews	0.94	1	
Poetry	0.76	0.83	1

	News	Reviews	Poetry
News	0		
Reviews	0.145	0	
Poetry	0.213	0.173	0

Table 3: Pairwise Spearman’s rank correlation coefficient r (top) and Jensen-Shannon distance (bottom) between the three data sets.

news – reviews – poetry. This result can possibly be attributed by the fact that poetry has more conceptually oral characteristics and is therefore closer to reviews than to news.

4.2 Evaluation of PoeTree.de

We compare the automatically created dependency relations that come with the PoeTree.de corpus with our manually created gold-standard annotations. As a preprocessing step, we again map the extended labels to the basic labels. Furthermore, we remove punctuation marks (i.e. tokens tagged as $\$.$, $\$($ or $\$.$ according to the STTS tagset) for two reasons: Firstly, the label punct with which they are attached to their heads is the most frequent one (see Table 2) but of little interest for the analysis of syntactic structure. Therefore, removing them makes the evaluation more informative. Secondly, for long sequences of parataxis, we systematically deviated from the UD guidelines for the label punct in that we attached it to the head of the last clause rather than the root node of the whole sentence to avoid overly long dependency arcs, thereby facilitating the manual annotation process.

UAS, LAS, CLAS For each poem, we calculate the following evaluation metrics: Unlabeled Attachment Score (UAS), which measures the percentage of tokens that are assigned the correct head; Labeled Attachment Score (LAS), which is the percentage of tokens that are assigned the correct head and the correct dependency label; Content-Word Labeled Attachment Score (CLAS, Zeman et al., 2017), which calculates LAS only for content words, ignoring function words. We use the evaluation script of the 2018 CoNLL shared task.¹³

¹³<https://universaldependencies.org/conll18/evaluation.html>.

Metric	PT.de	GSD	PT.cz	PDT
UAS	79.6 ± 8.7	82.8	85.0	95.0
LAS	68.9 ± 9.1	78.2	79.7	93.6
CLAS	59.2 ± 10.7	–	–	–

Table 4: Mean (and standard deviation) for different evaluation metrics. Column PT.de shows the results of our evaluation, whereas columns GSD, PT.cz and PDT show the results of evaluations from others (see the main text for details).

Table 4 shows the mean (and standard deviation) for each metric (see column PT.de), Figure 3 displays the distribution of scores across all poems. Overall, the results are rather poor. Straka (2018) reports clearly better results when applying UDPipe to the German-GSD corpus, which contains newspaper, reviews and web texts (see column GSD). The LAS scores in particular show a large gap.

Interestingly, when UDPipe is applied to a sample of 29 poems (6,591 tokens) from the Czech PoeTree subcorpus, it clearly outperforms both evaluations of German data, as reported in Plecháč et al. (2024) (see column PT.cz). However, the UDPipe parser seems in general to perform better on Czech data than on German data, as shown by column PDT in Table 4, which displays the evaluation results for UDPipe on the Prague Dependency Treebank (PDT), which consists of Czech newswire texts. Here the results are even higher.

The performance drop between news or web texts and other genres are in line with evaluation results e.g. on aesthetic writings (Salomoni, 2017) or fiction (Jelínek, 2017), not only for parsing but also for POS tagging (Haider, 2021) and other NLP tasks (see the overview in Bamman, 2020).¹⁴

The boxplot for UAS in Figure 3 shows that there is a very clear downward outlier, with UAS=53.8. This text also yields poor values for the other measures: LAS=50.0 and CLAS=37.8. Figure 4 shows an excerpt of this text with both annotations, the manual (top) and the automatic ones (bottom). The excerpt contains a very long coordination of adjectives, each of which is again modified by a preceding adverb. The head noun (*Poeten* ‘poets’) was

¹⁴The Arabic parser used by Al-Ghamdi et al. (2021) (see Section 1) achieves scores of UAS = 81.52 and LAS = 75.25. For the treebank of Spanish Sonnets, Ruiz Fabo et al. (2017) do not provide an evaluation of the automatic parses.

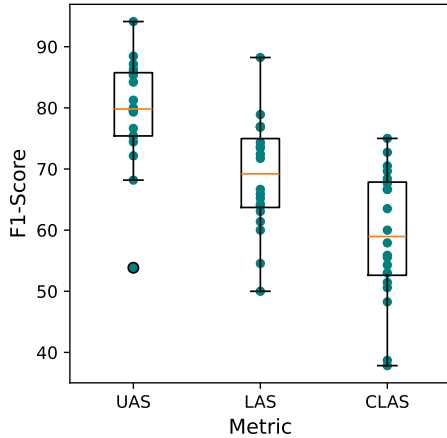


Figure 3: Distribution of UAS, LAS and CLAS scores.

Manual	PoeTree	F1	count
advcl	ccomp	0.39	7
parataxis	conj	0.28	26
obl	nmod	0.28	25
expl	obj	0.15	13
iobj	obj	0.13	11
nmod	obl	0.12	10
iobj	obl	0.11	7
expl	nsubj	0.09	12
root	appos	0.08	9
obl	obj	0.08	10

Table 5: Top ten most often confused dependency labels.

introduced in the previous line and is elided in the next lines. Therefore, in the manual annotation, the first of the adjectives (*begrabenen* ‘buried (ones)’) is treated as a substitute head, from which all further conj relations start. The system annotation, on the other hand, selects the last of the coordinated adjectives (*beglückenden* ‘enchanted (ones)’) as the head. This results in a large number of mismatches with regard to the structure and thus the low UAS value.¹⁵

Label confusion We use the F1 score as defined in Dipper et al. (2024) to rank the confusions between labels in a meaningful way, taking into account how often they were confused but also how

¹⁵In this excerpt, there are also two incorrect words: the forms *Langentzückten* and *Langbeglückten* – which do not exist in German – should actually be *klangentzückten* ‘sound-delighted’ and *sangbeglückten* ‘song-enchanted’. A translation of the excerpt is: ‘The exalted buried ones, and the striving living ones, the sensible ruling ones, the intimately unfolding ones, the lovingly shaping ones, the sound-delighted delightful ones, the song-enchanted enchanting ones, at experiences, at events’.

often they occurred overall.¹⁶ We only report confusions that occurred more than five times.

Table 5 shows the corresponding confusions. For example, seven expressions were manually annotated with *advcl* and automatically with *ccomp*. This distinction in particular also proved to be problematic in the study by Dipper et al. (2024), who annotated Middle High German texts. As we have seen, sentences are often strung together without an overt conjunction. Here it is often difficult to distinguish between pure parataxis and coordination without a conjunction. Other problematic labels concern the arguments of a verb, namely *obj*, *iobj*, *expl*. The confusion with *iobj* is partly due to the different criteria: while in the traditional UD-Treebanks for German *iobj* is only used for ditransitive verbs, the guidelines of Dipper et al. (2024) follow the suggestion of Zeman (2017), according to which *iobj* is used for all dative objects.

5 Conclusion

We presented an annotation study of German-language poetry of different centuries with UD-style syntactic relations. We showed that currently, the automatic analysis of poetry still seems problematic. At all levels – tokenization, sentence boundary determination, tagging, parsing – poems exhibit special properties that lead to faulty analyses. In this paper, however, only parsing was evaluated quantitatively.

One could argue that some or even many of the challenges discussed also apply to other non-standard language data, e.g. word repetition, anacoluthon and elision occur in spoken language, or deviations from standard spelling is typical of social media data. However, poets use these phenomena intentionally and deliberately in their poems, whereas in other data they often occur due to time pressure or attention deficits.

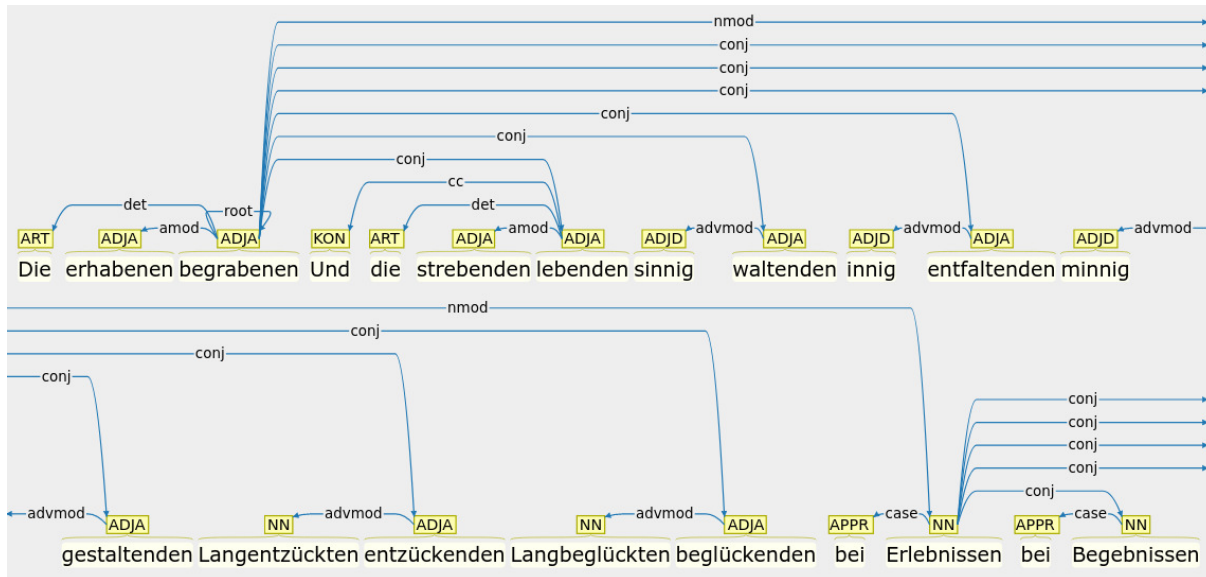
A major problem is that incorrectly determined sentence boundaries tear apart dependency-related phrases. In difficult passages, the correct syntactic relationships often only become clear on closer

¹⁶F1 is calculated as follows:

$$2 * \frac{a_1 l_1 * a_2 l_2}{a_1 l_1 + a_2 l_2}$$

with a_1 as the manual annotation and a_2 as the system annotation, and l_1, l_2 as the labels annotated by the human annotator and the system, respectively. Possible values are between 0 and 1, where 1 means perfect agreement if $l_1 = l_2$, and 0 means perfect disagreement if $l_1 \neq l_2$.

Manual annotation



System annotation

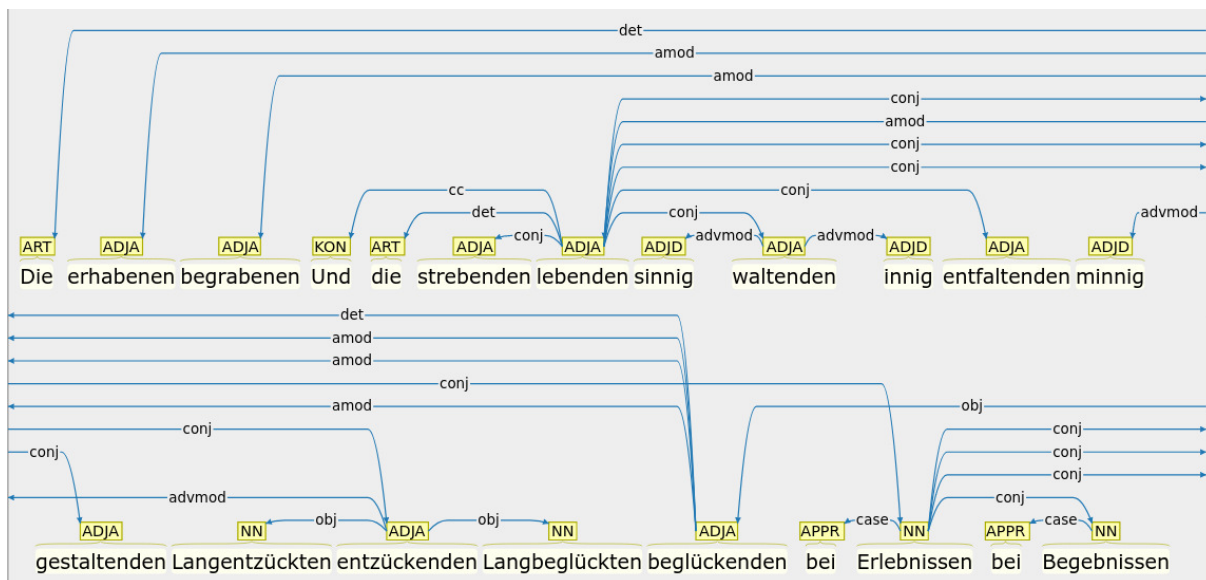


Figure 4: Excerpt of the outlier poem (ID 00-1734-0000-0003-7048-A#0u) with manual (top) and automatic annotations (bottom).

analysis and then often have a major influence on the interpretation. Therefore, we would like to propose that the sentence boundaries should be determined downstream and a dependency analysis should be carried out first, on the basis of which the sentence boundaries would then be determined (which of course would require a major modification of current dependency parsers). Alternatively, both tasks, parsing and sentence boundary detection, could be performed in parallel. In fact, this issue could concern other kinds of non-standard data as well, such as data produced by learners or children or in social media, which often does not adhere to standard punctuation rules.

Overall, we conclude that for studying syntactic properties of poems based on UD annotations, automatic parses are not yet reliable enough. Further manual analyses are necessary, also to provide training data for poetry-specific structures as reviewed in this paper.

Limitations

Our study is limited in that we have only annotated a small pilot corpus. We are planning to extend the data in the future. Furthermore, for the comparison of our manual annotations with automatically created annotations we have only used the automatically created dependency relations that come with the PoeTree.de corpus. The evaluation was end-to-end, i.e. we have not yet been able to evaluate the performance of a dependency parser that is based on, e.g., gold token and sentence boundaries.

Acknowledgments

We are very grateful to the anonymous reviewers for their constructive reviews.

References

- Sharefah Al-Ghamdi, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2021. [A dependency treebank for classical Arabic poetry](#). In *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*, pages 1–9, Sofia, Bulgaria. Association for Computational Linguistics.
- David Bamman. 2020. [LitBank: Born-literary natural language processing](#). In Jessica Marie Johnson, David Mimno, and Lauren Tilton, editors, *Computational Humanities, Debates in Digital Humanities*.
- David Bamman and Gregory Crane. 2011. [The Ancient Greek and Latin dependency treebanks](#). In Caroline Sporleder, Antal Van Den Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural*

Heritage, pages 79–98. Springer Berlin Heidelberg, Berlin, Heidelberg.

- Klemens Bobenhausen and Benjamin Hammerich. 2015. [Métrique littéraire, métrique linguistique et métrique algorithmique de l’allemand mises en jeu dans le programme *Metricalizer*](#). *Langages*, 199:67–87.
- Maria Cieschinger. 2016. *The Contraction of Preposition and Definite Article in German. Semantic and Pragmatic Constraints*. Ph.D. thesis, University of Osnabrück.
- Silvie Cinková, Petr Plecháč, and Martin Popel. 2024. [Rhymes and syntax: A morpho-syntactic analysis of Czech poetry](#). *Primerjalna književnost*, 47(2).
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Mirella De Sisto, Laura Hernández-Lorenzo, Javier De la Rosa, Salvador Ros, and Elena González-Blanco. 2024. [Understanding poetry using natural language processing tools: a survey](#). *Digital Scholarship in the Humanities*, 39(2):500–521.
- Matthew Timothy DeHass. 2024. [Parataxis in Latin colloquial and poetic texts : A treebank-based analysis](#). Master’s thesis, University of Missouri–Columbia.
- Eliane Delente and Richard Renault. 2017. [Projet Anamètre : présentation, limites et avancées](#). In *International Conference - Plotting poetry : on mechanically-enhanced reading*, number 7 in (Littératures), pages 73–92, Bâle, Switzerland. Presses Universitaires de Liège.
- Stefanie Dipper, Cora Haiber, Anna Maria Schröter, Alexandra Wiemann, and Maike Brinkschulte. 2024. [Universal Dependencies: Extensions for modern and historical German](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17101–17111, Torino, Italia.
- Stefan Grünewald and Annemarie Friedrich. 2020. [Unifying the treatment of preposition-determiner contractions in German Universal Dependencies treebanks](#). In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 94–98, Barcelona, Spain (Online). Association for Computational Linguistics.
- Thomas Haider. 2021. [Metrical tagging in the wild: Building and annotating poetry corpora with rhythmic features](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3715–3725. Association for Computational Linguistics.
- Thomas Haider. 2023. *A computational stylistics of poetry: distant reading and modeling of German and English verse*. Ph.D. thesis, University of Stuttgart.

- Hussein Hussein, Burkhard Meyer-Sickendiek, and Timo Baumann. 2018. [Automatic detection of enjambment in German readout poetry](#). In *Proceedings of Speech Prosody, 2018, Poznań*, pages 329–333.
- Tomáš Jelínek. 2017. [FicTree: A manually annotated treebank of Czech fiction](#). In *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017)*, pages 181–185.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.
- John Lee and Yin Hei Kong. 2012. [A dependency treebank of classical Chinese poems](#). In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 191–199, Montréal, Canada. Association for Computational Linguistics.
- Teresa Paccosi, Alessio Palmero Aprosio, and Sara Tonelli. 2023. Adding a novel Italian treebank of marked constructions to Universal Dependencies. *Italian Journal of Computational Linguistics*, 9(1).
- Susan Pintzuk and Plug Leendert. 2001. [The york-helsinki parsed corpus of old english poetry \(YCOEP\)](#). Oxford Text Archive.
- Petr Plecháč, Silvie Cinková, Robert Kolár, Artjoms Šeļa, Mirella De Sisto, Lara Nugues, Thomas Haider, and Neža Kočnik. 2024. [PoeTree: Poetry treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish](#). *Research Data Journal for the Humanities and Social Sciences*, pages 1–17.
- Petr Plecháč, Robert Kolár, Silvie Cinková, Artjoms Šeļa, Mirella De Sisto, Lara Nugues, Thomas Haider, N. Kočnik, Benjamin Nagy, Éliane Delente, Richard Renault, Klemens Bobenhausen, Benjamin Hammerich, Adiel Mittmann, Gábor Palkó, Péter Horváth, Borja Navarro Colorado, Pablo Ruiz Fabo, Helena Bermúdez Sabel, Kirill Korchagin, Vladimir Plungian, and Dmitri Sitchinava. 2023. [PoeTree. Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian and Spanish](#).
- Pablo Ruiz Fabo, Clara Martínez Cantón, Thierry Poibeau, and Elena González-Blanco. 2017. [Enjambment detection in a large diachronic corpus of Spanish sonnets](#). In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 27–32, Vancouver, Canada. Association for Computational Linguistics.
- Alessio Salomoni. 2017. [Toward a treebank collecting German aesthetic writings of the late 18th century](#). In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it)*, pages 292–297.
- Alessio Salomoni. 2021. [A UD Literary Treebank for German](#). PhD Thesis, Publisher: Università degli studi di Bergamo.
- Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. [Guidelines für das Tagging deutscher Textcorpora mit STTS \(Kleines und großes Tagset\)](#). Technical report, Universitäten Stuttgart und Tübingen, <http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf>.
- Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. [Building a user-generated content North-African Arabizi treebank: Tackling hell](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.
- Milan Straka. 2018. [UDPipe 2.0 prototype at CoNLL 2018 UD shared task](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.
- Daniel Zeman. 2017. [Core arguments in Universal Dependencies](#). In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 287–296, Pisa, Italy.
- Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Uřešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. [CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

PoeTree ID	Title	Author	Born	Created	Publ.	#Sents	#Tokens	#Stanzas	#Lines
dt.a.poem.18792	Mädchen .	Goethe, Johann Wolfgang von	1749	-	1819	2	25	1	4
dt.a.poem.10409	XXIV.	Heine, Heinrich	1797	-	1827	2	32	1	4
dt.a.poem.10555	ViI. Nachts in der Cajüte .	Heine, Heinrich	1797	-	1827	4	76	3	12
dt.a.poem.16173	227.	Rückert, Friedrich	1788	-	1837	6	132	6	12
dt.a.poem.3030	76.	Rückert, Friedrich	1788	-	1838	7	115	6	12
dt.a.poem.21583	Märznacht . 1884.	[anonymous]	-	-	1885	14	319	13	52
00-1734-0000-0003-28BA-1#0	135.	Hāfez, Šams o'd-din Moħammad ¹	1325	1357	1858	21	268	20	40
00-1734-0000-0003-2E39-A#0	10.	Hāfez, Šams o'd-din Moħammad ¹	1325	1357	1858	17	209	18	36
00-1734-0000-0003-F9D5-D#0	92.Schläge	Logau, Friedrich von	1604	1630	1872	4	70	1	6
00-1734-0000-0004-01CB-A#0	41. Grabschrift einer Buhlerin	Logau, Friedrich von	1604	1630	1872	2	24	1	2
00-1734-0000-0004-05D3-0#0	3.Franckenthal Friedens-Hindernuß	Logau, Friedrich von	1604	1630	1872	2	25	1	2
00-1734-0000-0003-576E-4#0	7. Was da braust	Herder, Johann Gottfried	1744	1773	1879	3	34	1	4
00-1734-0000-0005-8B68-7#0	MinenteFußnoten	Waiblinger, Wilhelm	1804	1817	1893	1	28	1	2
00-1734-0000-0003-7048-A#0	Breslauer Schillerfest	Hoffmann von Fallersleben, August H.	1798	1819	1841	2	65	1	17
00-1734-0000-0004-A660-A#0	Ausdruck der Empfindung	Rückert, Friedrich	1798	1827	1897	2	47	2	8
00-1734-0000-0002-EE1F-6#0	Begabung	Grillparzer, Franz	1791	1831	1960	1	20	1	4
00-1734-0000-0002-B719-B#0	Geduld!	Geibel, Emanuel	1815	1833	1918	15	184	6	24
00-1734-0000-0003-6F5D-5#0	Die Schlittenfahrt mit dem Schneemann	Hoffmann von Fallersleben, August H.	1798	1836	1976	9	97	6	12
00-1734-0000-0002-9E01-9#0	Nacht und Morgen	Eichrodt, Ludwig	1827	1841	1856	9	137	3	21
00-1734-0000-0002-577F-0#0	Es ist so still geworden ...	Conradi, Hermann	1862	1876	1911	23	255	5	40

Table 6: Overview of annotated poems. Titles are reported as specified in the JSON data of PoeTree.de. The columns ‘Born’, ‘Created’ and ‘Publ.’ refer to the year in which the author was born, the year the poem was created and the year it was published, respectively.

¹ Translated into German by Vincenz Ritter v. Rosenzweig- Schwannau, 1858.

Molyé: A Corpus-based Approach to Language Contact in Colonial France

Rasul Dent, Juliette Janès, Thibault Clérice, Pedro Ortiz Suarez, Benoît Sagot

Inria, Paris {firstname.lastname}@inria.fr

Abstract

Whether or not several Creole languages which developed during the early modern period can be considered genetic descendants of European languages has been the subject of intense debate. This is in large part due to the absence of evidence of intermediate forms. This work introduces a new open corpus, the Molyé corpus, which combines stereotypical representations of three kinds of language variation in Europe with early attestations of French-based Creole languages across a period of 400 years. It is intended to facilitate future research on the continuity between contact situations in Europe and Creolophone (former) colonies.

1 Introduction

Between the 15th and 19th centuries, several languages developed in colonized territories, which, while sharing a large amount of vocabulary with existing European languages, differ considerably in morphology and syntax. These languages are often labeled English-based [or lexified] Creoles, French-based Creoles, Portuguese-based Creoles, etc., according to the language they share most of their vocabulary with, which is itself called the lexifier. One long standing question has been why the grammars of these languages diverged from their lexifiers to a greater extent than the vocabulary (de Sousa et al., 2019). Much of the difficulty in answering this question stems from harsh social conditions discouraging linguistic documentation and environmental conditions destroying much of what had been documented (McWhorter, 2018).

For French-based Creole languages (FBCLs), which developed on islands and isolated continental settlements during the 17th and 18th centuries (Chaudenson, 2001)¹, reliable documentation largely dates from the mid-late 18th century onward (Hazaël-Massieux, 2008). However, we note that the formative period of FBCLs coincided

with a period of French political and cultural dominance and extensive literary production known as the Grand Siècle. The cultural works of the period are replete with numerous stereotypes of the speech of several social groups, such as urbanized peasants and Swiss soldiers. Despite various issues detailed by Ayres-Bennett (2000), these representations are relevant for FBCLs insofar as they demonstrate several interesting morphosyntactic developments.

Here, we introduce the Molyé corpus, which regroups 68 works that contain examples of either the aforementioned stereotypes or early attestations of FBCLs. This list has been curated from a larger collection of 301 documents identified at the time of publication.² We begin by giving an overview of related corpora and how we approach historical linguistics as an instance of multi-label language identification. After giving some linguistic context, we also explain the process of identifying Creole-like features in French literary works, encoding said works into XML-TEI, and then compiling groups of quotes into a timeline. Finally, we present summary statistics and conclude by giving examples of how our corpus highlights intra-European contact.

2 Related Work

In recent years, Creole languages have garnered attention in the field of natural language processing. On the one hand, Lent et al. (2022b) have explored how these languages challenge the assumed desirability of certain applications. On the other hand, Lent et al. (2022a) and Robinson et al. (2023) argue that language models for concrete problems may shed light on theoretical issues as well. Simultaneously, in computational historical linguistics, List (2024) has declared the inferral of morpheme boundaries and the detection of layers of language contact to be major open problems. Our work addresses both the paucity of early Cre-

¹Except Tayo in 19th century New Caledonia.

²The corpus can be accessed and downloaded at the following address: <https://github.com/defi-colaf/Molye>.

ole documentation and the issue of multiple layers of language contact through the applied lens of language identification.

2.1 (Digital) Diachronic Corpora

For several Creolophone regions, such as Louisiana (Neumann-Holzschuh, 1987), the Caribbean (Hazaël-Massieux, 2008), Réunion (Chaudenson, 1981), and Mauritius (Baker et al., 2007; Chaudenson, 1981), diachronic corpora have been compiled in print. However, to our knowledge, only the Mauritian corpus has been systematically digitized and made readily accessible (Fon Sing, 2013). Beyond this, certain historical works have been digitized for inclusion in analysis-oriented private diachronic corpora (Mayeux, 2019), or for applied goals like machine translation (Robinson et al., 2024), and others have been individually published by groups such as the European Group for Research on Creole Languages (Hazaël-Massieux, 2013a,b).

To digitize documents in a way that can facilitate reuse, we rely on the standards of Text Encoding Initiative (TEI) (TEI Consortium eds., 2023). Adherence to these guidelines has produced diachronic corpora which span several centuries, such as Favaro et al. (2022). For the languages of France, Bermudez Sabel et al. (2022) have addressed some of the challenges of building comparable corpora for the parent-daughter pair of Latin and French. Similarly, Ruiz Fabo et al. (2020) explore how digitizing 19th century Alsatian theatre aids sociolinguistic studies.

2.2 Multi-Label Language Identification

Algorithms for determining the language of a given text generally rely on tokenizing the text and comparing the tokens against a learned (or explicitly defined) representation of a language (Jauhiainen et al., 2019). For analytic languages written in the Latin alphabet (i.e. FBCLs), tokens generally align with either words or letters. With closely-related languages, there is sometimes only a difference of a singular word or even letter between one variety and another, even in longer documents (Ljubesic et al., 2007; Caswell et al., 2020). In these cases, we can specify disjunctive features such as words/phrases that are thought to separate the varieties to either affirm or reject a label. In the absence of such features, the same string may be valid in multiple languages, which can make it more accurate to assign multiple language labels

to the same string (Bernier-Colborne et al., 2023; Keleg and Magdy, 2023).

3 Linguistic Background

The backbone of our corpus is applying multi-label language identification based on disjunctive features across time. In concrete terms, we sought out distinctly “Creole” features in Europe before and during the colonial expansion. As such, we briefly review a few characteristics of FBCLs, followed by French literary stereotypes.

3.1 French-based Creole Languages

3.1.1 Description

While the notion that all Creoles can be defined in purely linguistic terms, as explored by McWhorter (1998); Bakker et al. (2011), is controversial, FBCLs are agreed to share several traits which distinguish them from standard French. Firstly, they generalized the use of tonic pronouns in places where the latter would use weak clitic pronouns (Syea, 2017). In cases where French does not have a weak pronoun (i.e. ‘nous’), they still differ by not allowing preverbal cliticization of object pronouns. Additionally, while French relies on a system of fusional conjugations, where verb endings mark person, number, tense, aspect and in the case of the past participle, gender, at the same time, FBCLs add person-invariant combinations of Tense-Aspect-Mood (TAM) markers (Syea, 2017; Baker and Corne, 1982). These differences are demonstrated by the anteriority marker ‘té’, and the conditional marker ‘sré’ in the phrase ‘Pour sûr si vou **té** capab changé vou lapo pou so kenne, vou **sré** pa di non’ (Mercier, 1881)³. Furthermore, FBCLs do not have an explicit copula in several structures where one is required in French (and English), as demonstrated by the phrases ‘Comme **vous bel**’⁴ and ‘vou **papa riche**’⁵ in Figure 2.

3.1.2 Theories of Origins

As previously stated, the relationship of Creole languages to lexifiers remains a topic of intense debate. For this work, one relevant hypothesis, as explored by Chaudenson (2001), suggests that the accumulation of the defining characteristics occurred over several waves of second language acquisition, as opposed to being the result of a complete break in

³Surely, if you could trade your skin for his/hers, you would not say no.

⁴how **you** [are] **beautiful**.

⁵you[r] **dad** [is] **rich**.

transmission of syntax, as suggested by McWhorter (2018) and Thomason and Kaufman (1988). Another line of inquiry explores the extent to which “foreigner talk”, which is to say a particular kind of simplified register that people adopt when they feel their interlocutors do not have sufficient competence in the language, may have contributed to certain developments in Creole morphology and syntax (Ferguson, 1981, 1975). For Portuguese- and Spanish-based Creoles, there is a long history of triangulating Iberian versions of foreigner talk with early modern literary stereotypes and contemporary Afro-Hispanic varieties to get an idea of the range of linguistic variation in the early modern Iberian empires (Kihm, 2018; Lipski, 2001). In the following section, we explore how a similar approach can be applied to French.

3.2 French Literary Stereotypes

Up to the 20th century, most people in France spoke regional languages (Lodge, 2003). In the Northern half of mainland France, most of these languages are part of the Oïl dialect continuum, which is itself part of a larger Western Romance continuum. However, non-Romance languages such as Breton (North-West) and Flemish (North) are spoken as well. From the Middle Ages on a particular Oïl variety, associated with prestigious actors was gradually codified into the standard language of the Kingdom of France. This variety was also adopted as a *lingua franca* throughout Europe, as an alternative to Latin. During the 17th and 18th centuries, the process of codification culminated in a well delimited variety known as Classical French.

However, the codified “bon usage”, was not the only supralocal speech used in France. Even within the Paris region, there was a great deal of variation within what could be considered “French” (Wittmann, 1995). In broad terms, we distinguish three types of variation: dialectal and sociolectal variation from the Oïl domain, standard French with regional accents, and interlanguages, especially from L1 speakers of Germanic languages⁶. In all three of these cases, we find stereotyped combinations of a finite number of highly stigmatized features in a variety of works, including plays, novels, songs, and personal letters.

⁶Other phenomena, such as the mix of various forms of Occitan in *Monsieur de Pourceaugnac* described by Sauzet and Brun-Trigaud (2015), are beyond our immediate scope.

3.2.1 Peasant French

By the early 1600s, several features of rural usage in the outskirts of Paris (and Western France), such as the combination of clitic pronoun ‘je’ with the plural affix ‘-ons’, were developed into a convention for representing lower class characters in literature (Lodge, 1991; Ayres-Bennett, 2004), as seen in this example from *La Mère confidente* (Marivaux, 1735): ‘Je savons bien ce que c’est; j’ons la pareille.’⁷ Although this stereotype was frozen relatively early on, the highlighted combination was used in France and its colonies throughout the colonial period and still exists in Acadian French in particular, albeit more commonly as a plural form (King et al., 2004).

3.2.2 Gascon Accent

French also came to be spoken as a second language in areas where the regional languages were even more different from French. In these cases, the native languages had some influence on pronunciation. In classical French theatre, one common stereotype of such regional pronunciation is the Gascon accent, which can be identified through its betacism (conflating b and v) and fronting of the schwa (replacing e with é). The character Fontignac from *L’île de la raison* (Marivaux, 1727) demonstrates the convention with this line: ‘...bous mé demandez cé qué bous êtes ; mais jé né bous bois pas ; mettez-bous dans un microscope.’⁸

3.2.3 Germanic Baragouin

Germanic Baragouin⁹ (henceforth just Baragouin) is our name for a group of stereotypes which simultaneously combine traits of foreigner talk, foreign accents, and Oïl dia- and sociolectal variation. In the early modern period, there are two main variations: the Anglo-Baragouin attributed to English (and Scots) speakers, and Continental Baragouin associated with German and Dutch, and more specifically, Swiss and Flemish speakers (Leach, 2020; Damm, 1911). A third, industrial-era Flemish Baragouin also developed around the turn of the 20th century in the cities of Tourcoing and Roubaix near the French-Belgian border (Landrecies, 2001). The main differences between

⁷We/I know what [the task] is, we/I have a similar one.

⁸... You ask me what you are; yet I do not see you. Put yourself in a microscope.

⁹The word “baragouin” [gibberish] was also used to describe a variety of contact phenomena ranging from accented pronunciation to genuine pidgins like that used with the Caribs in the Lesser Antilles (Wylie, 1995).

these sub-groups of Baragouin lie in phonetics. The Continental Baragouin generalizes final-consonant devoicing into a complete neutralization of several consonant pairs, such as /b/-/p/, /k/-/g/, /v/-/f/ and /t/-/d/. Similarly, the industrial-era Flemish version features palatal fronting of /ʃ/ and /ʒ/ to /s/ and /z/. These traits are mostly absent in the English version.

In terms of morphosyntax, Baragouin shares some traits with Creoles, such as the generalization of strong pronouns, weakening of grammatical gender, and reduced verbal inflection (Haas, 2015). However, Baragouin also retains an overt copula and systematically inserts third-person pronouns before verbs, which results in sentences such as ‘**Toi li être**, par mon foi, la plus pelle meilleure humeur du monde ¹⁰’ (Guelette, 1740). The latter features have a special importance, which we explore further in Section 7.1.

4 Corpus Creation

The compilation of the corpus was realized in three overlapping phases. During the first phase, we identified documents which contained n-grams thought to be highly disjunctive between French and various FBCLs. After identifying the documents the next step was to convert them relevant samples into the XML-TEI schema of a broader project. Lastly, we classified the documents by location and period and extracted the relevant quotes into a combined XML document to facilitate the preliminary analysis presented in Section 6.

4.1 Document identification

The basic strategy was to search Gallica, the digitized library of the Bibliothèque nationale de France ¹¹, Delpher, its Dutch equivalent, and later Google Books for disjunctive n-grams. Examples include monograms (e.g. ‘mo’, ‘to’, ‘yé’), bigrams, e.g. (‘mo(n) femme’, ‘mo(n) z’ enfant’), and higher n-grams. Due to variation in both French orthography and the conventions/contact varieties themselves, an iterative approach was taken, with documents collected on the first pass providing more “unusual” n-grams for subsequent searches. In the earliest stages, we did not note the exact searches, but later began to record the search terms as well. In a later stage, we also added several Cre-

¹⁰By my faith, you are [lit.**you it be**] the most beautiful best humour of the world

¹¹National Library of France

ole sources known through secondary literature in order to facilitate in-depth diachronic comparison.

Because we are working with stereotypes, a certain level of similarity was to be expected. Nevertheless, in some cases, we found that certain works go into the realm of explicit reference and/or pastiche. For direct quotation, there is *Les fêtes de l’amour et de Bacchus* which includes a reprise of the linguistic humor from *Le Bourgeois Gentilhomme*, among other pieces. As far as pastiche, we can highlight the early 16th century *Testament du Gentil Cossoys* and its early 17th century reprise, the *Testament d’vn Escossois*. The latter is a simultaneously condensed and updated version of the former. Thus where the original reads ‘Adiou par tout noble royaulm de Frans / Adiou comman le povre pals de Cos...’¹² (Smith, 1920), the reprisal has ‘Ady par tout le Royaume de France/ Premiere-ment ady le pay de Coss...’¹³ (Sigogne, 1620)

Search	Lang Type	Document	Year
“ly va”	Baragouin	Francion	1630
li-même	Peasant	L’Épreuve	1740
conné li	L. Creole	L’autre monde	1855

Table 1: Sample Searches and Documents

4.2 Encoding Documents

Given both the large number of documents it was necessary to establish an order of priority for incorporating works into the corpus. We initially focused on both Baragouin and Peasant French in works of classical theatre that had already encoded by sources such as theatre-classique.fr (Fièvre, 2007). Beyond the core of classical French theatre, however, a wide variety of genres are represented. These include poetry, songs, religious material, short prose, and an entire novel. The subject matter exhibits a similar degree of variability. In the Baragouin section alone, we find, among other things, two mock-testaments, a criticism of military leadership, a love letter, and a discussion about the political implications of an ongoing civil war.

After treating the extant XML, we explored semi-automatic generation of XML-TEI documents from semi-structured sources such as Wikisource, as

¹²Adieu to all noble kingdom of France / Adieu likewise poor Scotland

¹³Adieu to all the Kingdom of France/ Firstly adieu Scotland

well as directly from scanned documents. In the former case, we used relatively simple custom Python scripts to facilitate conversion to TEI, such as wrapping all of the lines in a `<p>` (paragraph) or `<l>` (line/verse) tag, and then identifying divisions and headers manually. In the latter case, this involved a considerable amount of manual transcription due to the diversity of genres and formats. For shorter works, such as poems and songs, we used eScriptorium (Kießling et al., 2019) to perform text recognition with the CATMuS Print model (Gabay et al., 2024). However, more complex layout (e.g. newspaper) were transcribed manually. For longer works, we entered the relevant quotes directly into a file of excerpts.

4.3 Linguistic Annotation

Since this corpus is in large part intended to illustrate a sociolinguistic continuum assigning discrete linguistic labels poses distinct challenges. Although it is clearly anachronistic to speak of “[Colony] French/Creole” before the founding of a given colony, we observe that in certain cases, namely in Réunion and Louisiana, the “approximative French”, “pidginized French”, or “pre-Creole” (depending on one’s point of view) bears striking continuity with Baragouin at the morphological and syntactic levels. In a parallel fashion, early texts which are clearly “Creole”, such as “La passion de Notre Seigneur selon St Jean en Langage Negre”, display combinations of features which make it difficult to say *which* Creole based on purely linguistic data.

Following the brief outline given in 3, we distinguish between five main kinds of language: Classical French (met-fr), Peasant French (fra-dia), (Gascon) Accented French (fra-gsc), Baragouin (subdivided into fra-ang, fra-deu, and fra-nld), and (pre-)Creoles. The Creole portion is in turn subdivided into four regions and labelled using the respective ISO codes: Réunion (rcf), Louisiana (lou), Haitian (hat), and French Guianese (gcr). For the initial work, we have somewhat simplified the question of diachronic and dialectal continua by assigning one label based on the territory a document claims (or has been presumed) to represent, with the exception of grouping the earlier “Flemish” baragouin with the German one rather than the later Flemish Baragouin, based on the differences described in Section 3.2.3.

For adding linguistic labels to documents, we

```
<div type="scene" n="10">
...
<sp who="JACQUES" xml:lang="mau">
  <speaker>JACQUES.</speaker>
  <p>... Enfin pourtant , li jetté son zépée ,
    li remetté pistolet dans son place ,
    li prendre son plume , li assisé tranquille ,
    et li fini écrire sa billet là moi porté vous.
    Ah vlà li.
  </p>
</sp>
<sp who="STRAFFORD" xml:lang="fra-ang">
  <speaker>STRAFFORD lit le billet haut.</speaker>
  <p>» Vous avez raison , monsieur ,
    je suis mort pour vous et pour votre ami » .
  <stage> ( Il parle. )</stage>
  Toi voir lui mort [etc...]
</p>
</sp>
<sp who="BELTON" xml:lang="met-fr">
  <speaker>BELTON.</speaker>
  <p>Moins que jamais ;
    c'est absolument une énigme pour moi.</p>
</sp>
</div>
```

Figure 1: This excerpt from Scene 10 of *Le duel singulier* (Dorvigny, 1800) shows how we tag language usage by speaker. It includes standard French alongside Anglo-Baragouin and an unspecified Creole with Mauritian characteristics. [formatting adjusted]

used two complementary rule-based strategies. For plays where one character (or more) uses non-standard speech throughout, we simply identified the `<sp>` (speech) tags associated with that character and inserted an `xml:lang` attribute with the corresponding label, which allowed us to keep associations between characters and speech turns. Additionally, we added tags at the `<p>` level to facilitate text extraction.

For prose, keeping track of specific characters was more difficult. Initially, we tried implementing key-ngram-based regex patterns. Because our languages of interest are frequently embedded in longer French passages, a preprocessing step of sentence tokenization was implemented. Although our disjunctive n-grams generally correspond to words, we use character-level regex patterns that incorporate a special boundary symbol to minimize multi-level tokenization. For the initial annotation, the presence of any one disjunctive n-gram was sufficient to trigger the relevant label. While this method was very useful for highlighting interesting passages, manual retouching was necessary to fix issues of imperfect sentence tokenization, as well as missed examples. In Figure 2, we find a reported clause in Louisiana Creole that is not marked because it contains no disjunctive words, followed by a reporting clause in French¹⁴, that is unintentionally included with correctly identified

¹⁴dit l’esclave d’une voix caressante’ [said the slave with an affectionate voice].

Target/Region	Label	Works	Tokens	Timespan
Normative French	met-fr	35	37066	1649-1779
Peasant	fra-dia	14	27825	1665-1740
Gascon	fra-gsc	4	4530	1672-1800
Anglophone	fra-ang	4	4441	1509-1800
Continental Germanic	fra-deu	25	6899	1580s~1779
Flemish (Tourcoing/Lille)	fra-nld	4	2664	1880-1932
Réunion	rcf	3	10713	1760s, 1830s
Lesser Antilles (Martinique)	gef	2	477	1671
Haiti	hat	4	7395	1730s~1802
Louisiana	lou	10	26068	1748-1895
French Guiana	ger	2	43414	1796, 1885
Mauritius (tentative)	mau	1	196	1800

Table 2: An overview of the linguistic and temporal spread of the corpus.

Creole speech in the following sentence. The third sentence is marked as expected.

```
<p>
« Comme vous bel !
<s xml:lang="lou"> dit l'esclave d'une voix caressante ;
vou gagnin ain ti lair si tan comifo ! </s>
<s xml:lang="lou">vou popa riche, mo sûr ;
di li achte moin.</s>
...
</p>
```

Figure 2: Uncorrected semi-automatic annotation of *L'Habitation Saint-Ybars* (Mercier, 1881)

4.4 Compiling Extracts

After adding language tags at the document level, we created a composite timeline that balances facilitating direct comparison between excerpts with giving some level of contextualization. For plays, we extracted scenes where at least one of the `<sp>` turns contained an `xml:lang` attribute with an appropriate value, as demonstrated by Figure 1. By extracting the entire scene, we include samples of normative French and retain the coherence of the conversation to some extent. For monolingual poems, we included the entire poem, albeit possibly excluding meta-linguistic commentary such as notes. For prose, we implemented a multi-level extra process of first trying to identify broad tags like `<p>` based on the `xml:lang` attribute, and then narrower tags like `<s>` only if they were not already included as part of a broader group. In Figure 2, the overall paragraph would be assumed to be French, so only the lines within the `<s>` tags would be extracted, which is why correcting the linguistic annotation is important.

4.5 Balancing

As exemplified by the Gascon accent, the literary conventions can be summarized using a relatively short list of rules. This means that there is a degree

of diminishing returns to adding additional examples once we have a basic understanding of said rules. As such, we did not concern ourselves with attempting to create a statistically balanced corpus. In particular, due to the more labor-intensive nature of (semi)-manual encoding, we deprioritized the Peasant French variety early on because it has already received more careful study, and instead focused on the earliest and latest attestations of Baragouin. This may create the impression that literary Peasant French was primarily a 17th century phenomenon. However, this stereotype remained in use until the 19th century. Along similar lines, we did not include many attestations of Mauritian Creole precisely because a digitally accessible diachronic corpus to the same effect already exists (Fon Sing, 2013; Baker et al., 2007).

5 Corpus Presentation

Overall, we found 301 historical works which demonstrate features relevant for the history of FB-CLs. We have selected excerpts from 68 of these works to form the basis of the first version of the corpus. The earliest text is “Le Testament du Gentil Cossoys”, written anonymously around 1509, and the most recent is Jules Watteuw’s “Belle Réponse”, published in 1932. The main corpus consists of a single, publicly available XML file containing bibliographic information for the collection, followed by a body which contains “TEI” tags that regroup the relevant selections from each work and are accompanied by their own brief bibliography section. From this file, one can create customized subcorpora that correspond to specific questions by specifying a date range and the language labels that are to be considered.

At present, the corpus contains a total of 188,866 tokens (whitespace tokenization), excluding meta-

Target/Region	Infinitive	Inflected	TAM	CE	Tokens
Normative	105	1328	129	254	37066
Peasant	76	1006	129	251	27825
Gascon	14	131	16	47	4530
Anglophone	74	32	7	5	4441
Continental Germanic	89	62	11	13	6899
Industrial Flemish	0	44	0	18	2664
Réunion	5	125	54	2	10713
Haiti	0	157	102	27	7395
Louisiana	10	1086	944	129	26068
French Guiana	1	1001	950	40	43414

Table 3: Attestations of different forms of ‘être’. TAM and CE cover creolized inflection.

data. Because of the historical focus of the text, all of the primary sources are in the public domain, and most are readily consultable online. In these cases, we also retain cached copies with additional bibliographic information. In the cases where quotes have been included from printed secondary sources, we do not include metalinguistic commentary. Table 2 provides a high-level summary of the varieties we distinguish and their relative sizes and time spans.

6 Preliminary Results

Since the main effort of this work has consisted of gathering and grouping multiple non-standardized varieties, proceeding directly to quantitative methods presents special challenges. For the initial demonstration, we provide a few qualitative observations and show how we can support them through relatively simple frequency-based methods, with a particular focus on the relevance of Baragouin¹⁵.

6.1 First Person Pronoun: Mo(è)

During the colonial era, the French pronoun ‘moi’ [me] had two primary variants: **mwe** and **mwa**. FBCLs can be grouped according to which form of ‘moi’ became the subject pronoun. The first group, consisting of Haitian and Lesser Antillean Creoles, predominantly uses **mwě**, which is clearly a nasalized version of **mwe** (Hull, 1979). The second group, comprised of Mauritian, Seychellois, French Guianese, and Louisiana Creoles, uses the form **mo**. This division corresponds to further differences in the pronominal system, with the first group also using case-invariant pronouns and marking possession through postposition, while the second group distinguishes between subject and oblique variants

and uses proposed possessive adjectives¹⁶. Although **mo** is tied to **mwa**, its exact origins are less clear. Furthermore, there is documentation that **mo** was once used by the first group, before being replaced in the 1900s (Hazaël-Massieux, 2008).

Several of our documents shed new light on the relationship between these two variants. Firstly, beyond the canonical **mo**, we also found examples of ‘moué’, ‘moé’, ‘moè’, ‘moë’, and ‘moa’ in 19th century Louisiana alone. In Jobey (1860, p.189), for example, includes ‘*Moè té cré bien, moè perdu papier la yest*’¹⁷, which combines the Caribbean-like **mwe** with the Louisiana-specific definite plural marker **laje** (spelled ‘la yest’). By itself, this can be explained by 19th century New Orleans’ status as a crossroads of French- and Creole-speaking networks. Secondly, however, we found numerous attestations of **mo** in Flemish Baragouin. For example, the opening line of “Poutche” (Watteuw, 1927) is ‘Accoute un fos, **mo** ne pas bête’¹⁸. The latter may help explain **mo** as one innovation which diffused from Europe alongside **mwe**, rather than a parallel innovation.

6.2 Copula: ê(tre)

Additionally, we noticed that Baragouin has a tendency to overuse the infinitive ‘être’ (to be), rather than either conjugating the verb like French, or omitting the copula as in FBCLs. We began quantifying this variation by measuring the frequency of two basic patterns: the infinitive, and all inflected forms. We further tracked two subsets of inflected forms that have been integrated into various FBCLs: (precursors of) TMA markers ((**e**)**te**), **s(r)e**, **s(r)a** and orthographic variants thereof),

¹⁵The following section uses broad IPA in bold.

¹⁶Exceptionally, Réunion uses **mwě** with case distinctions.

¹⁷I had really though I lost the papers.

¹⁸Listen up, I ain’t [lit. me not] stupid.

as well fusions involving the pronoun ‘ce’. For demonstrative purposes, we set aside the samples for the Lesser Antilles and Mauritius, since they are particularly limited. Unfortunately, we could not take into account the clause-restricted copula **je** due to it being homophonous with the much more frequent third-person plural pronoun and a derived plural marker in Louisiana and French Guiana.

Table 3 demonstrates the results of this experiment. As expected, Normative French, Peasant French, and Gascon-accented French all use a wide variety of inflections. In contrast, the FBCLs the Americas retain specific grammaticalized uses, such that ‘être’ is rare, while inflected forms largely correspond to either TAM markers or presentatives with ‘ce’¹⁹. Réunion, which is distinguished among FBCLs for retaining French auxiliaries, stands out as transitional. In contrast to both groups, both Anglophone and Continental Baragouin (but notably not later Flemish) generalize use infinitive ‘être’ more than inflected forms, but do not completely discard the latter.

7 Discussion

7.1 Missing (L)(i)nks

By itself, the generalization of ‘être’ shows that decreased use of inflection and copula deletion, two traits of FBCLs suggested to indicate pidgin origins by McWhorter (2018), did not necessarily develop at the same time nor for the same reason. Beyond this, however, we are able to directly tie one process underlying the generation of Baragouin to one Creole language in particular: Réunion Creole (RC).

As Hull (1993, p.393) observes, the subject pronoun **li**, shared by all FBCLs, is employed by a Swiss German in *Le Bourgeois gentilhomme* in place of ‘il’. More specifically, as Damm (1911) remarks, the systematic insertion of this third-person pronoun before verbs, as mentioned in Section 3.2.3, is particularly reminiscent of RC. In early texts demonstrating a transitional variety between French and RC, we find sentences such as ‘**Moi i crois** vrai, bien vrai dans mon cœur n’en a bon Dieu’²⁰ (Bollée, 2007). In both Baragouin and RC, this preverbal pronoun also fuses with auxiliaries, as in this example from *Les filles errantes* (Regnard, 1690): ‘**Moi l’être** un étranger qui cher-

chir à logir dans sti ville.’²¹ and the Réunionese ‘**Moi l’est** bien content voir à vous’²² (Héry, 1883).

Although the exact function and source of the preverbal marker in Réunionese Creole are both debated, one common interpretation is that it marks finiteness on verbs and originated as a generalization of third person reprise pronouns (Bollée, 2007). Interestingly, a similar generalization of third-person **object** pronouns is observed in Spanish-language representations of Africans as early as the 17th century, and comparable phenomena continue to exist in varieties of Spanish in the Americas influenced by Quechua and Nahuatl (Lipski, 2001). In our corpus, we also observe that ‘li’ in particular also appears in Peasant French, primarily as a clitic indirect object. As Baragouin also inserts preverbal pronouns in sentences that use the French first-person subject clitic ‘je’, the inserted preverbal pronoun corresponds to a few homophonous French subject, object, and adverbial pronouns. This in turn suggests our data is relevant for contact scenarios beyond FBCLs.

7.2 The Bigger Picture

Beyond tracking individual features, our corpus offers a window into the broader sociolinguistic context of French in the early modern period. In the case of the first person pronoun, despite the temporal mismatch, the specificity of ‘mo’ points to the Low Countries as a point of interest. Upon closer examination, several works spell out a network connecting Swiss soldiers to this region and Paris in the context of the French-Hapsburg wars such as a 1692 “Air suisse ou flamand” which references the Nine Years’ War in Mons, Namur and Maastricht directly. This detail is of interest for Louisiana and Mauritius, where German-speaking settlers and soldiers played important roles in the French colonization in the 1720s. (Vaughan, 2005; Klingler, 2003; Baker and Corne, 1982).

Along similar lines, *Le duel singulier* stands out as a ready-made case study. This play combines normative French, the Gascon accent, Anglo-Baragouin, and an unspecified Creole, as exemplified in Figure 1. As such, it bolsters theories that the FBCLs of the Caribbean region may have developed during the period of Anglo-French cooperation during the early 17th century on islands such as Saint-Christophe and Tortuga (Parkvall, 1995). Fur-

¹⁹And **je** which we left out as explained above.

²⁰I [lit. me **it**] believe true, very true in my heart there is [good] God.

²¹I am [lit. me **it be**] a foreigner looking to lodge in this town

²²I am [lit. me **it’s**] glad to see you.

thermore, the Baragouin can be cross-referenced against the Law French of English courts of that period (Löfstedt, 2014).

8 Conclusion

In short, we have introduced the Molyé corpus, a new resource which puts French literary stereotypes alongside early forms of several French-based Creole languages. We have shown that restructuring of the French pronominal and verbal systems are attested throughout the 16th, 17th, and 18th centuries, and specifically associated with speakers of Germanic languages. Although stereotypes like the conventionalized Baragouin only address a fraction of the real linguistic variation of the period, our corpus nevertheless raises important questions about how people communicated in lands where French and Germanic languages came into contact. Furthermore, it shows that at least some of the divergences between FBCLs and French can be traced back to developments which were already underway in Europe.

Limitations

The major constraint of this work has been converting unstructured works into XML-TEI. As mentioned in the methodology, this involved complete re-transcription in some cases. Overall, we found more than 200 pertinent documents, but were only able to include one third of them. In particular, we had to leave out works in regional languages of France such as Picard, Walloon, and Poitevin. Similarly, we did not address some relevant phenomena, such as the 17th century Carib Baragouin and the 19th century Tirailleur French in order to maintain the scope of the work. Although we are well aware of such varieties, we found few instances using our method, and thus leave them as natural targets for future work.

Ethics Statement

The main idea of this article is that European literary stereotypes from before and during the colonial period can help fill in the some gaps in the early history of (French-based) Creole languages. As such, many of the primary and secondary sources that we have compiled contain negative imagery and commentary regarding various social groups. Sharing such sources should not be taken as endorsement of the views contained therein.

Acknowledgements

This work was primarily funded by the Inria “Défi”-type project COLaF. This work was also partly funded by the last author’s chair in the PRAIRIE institute funded by the French national agency ANR as part of the “Investissements d’avenir” programme under the reference ANR-19-P3IA-0001.

References

- Wendy Ayres-Bennett. 2000. *Voices From The Past: Sources of Seventeenth-Century Spoken French*. *Romanische Forschungen*, 112(3):323–348.
- Wendy Ayres-Bennett. 2004. *Sociolinguistic Variation in Seventeenth-Century France: Methodology and Case Studies*. Cambridge University Press.
- Philip Baker and Chris Corne. 1982. *Isle de France Creole: Affinities and Origins*. Karoma, Ann Arbor, Mich.
- Philip Baker, Guillaume Fon Sing, and Vinesh Y. Hookoomsing. 2007. The corpus of Mauritian Creole texts. *The making of Mauritian Creole. Analyses diachroniques à partir des textes anciens*, (9):1–61.
- Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. *Creoles are typologically distinct from non-creoles*. *Journal of Pidgin and Creole Languages*, 26(1):5–42.
- Helena Bermudez Sabel, Francesca Dell’Oro, Cyrielle Montrichard, and Corinne Rossari. 2022. *Setting Up Bilingual Comparable Corpora with Non-Contemporary Languages*. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 56–60, Marseille, France. European Language Resources Association.
- Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. *Dialect and Variant Identification as a Multi-Label Classification Task: A Proposal Based on Near-Duplicate Analysis*. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.
- Annegret Bollée. 2007. Deux textes religieux de Bourbon du 18e siècle et l’histoire du créole réunionnais.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. *Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Robert Chaudenson. 1981. Textes créoles anciens: La Réunion et île Maurice: Comparaison et essai d’analyse. *Kreolische Bibliothek*.

- Robert Chaudenson. 2001. *Creolization of Language and Culture*. Routledge.
- Otto Damm. 1911. *Der deutsch-französische Jargon in der schönen französischen Literatur*. Emli Eberling.
- Silvio Moreira de Sousa, Johannes Mücke, and Philipp Krämer. 2019. *A History of Creole Studies*. In *Oxford Research Encyclopedia of Linguistics*.
- Louis Francois Dorvigny. 1800. *Le duel singulier, comédie en un acte et en prose*.
- Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi, and Simonetta Montemagni. 2022. Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 94–100, Marseille, France. European Language Resources Association.
- Charles A. Ferguson. 1975. *Toward a Characterization of English Foreigner Talk*. *Anthropological Linguistics*, 17(1):1–14.
- Charles A. Ferguson. 1981. ‘Foreigner Talk’ as the Name of a Simplified Register. *International Journal of the Sociology of Language*, 1981(28):9–18.
- Paul Fièvre. 2007. *Théâtre classique*.
- Guillaume Fon Sing. 2013. *Corpus de textes anciens en créole mauricien*.
- Simon Gabay, Thibault Clérice, Pauline Jacsont, Elina Leblanc, Marie Jeannot-Tirole, Sonia Solfrini, Sophie Dolto, Floriane Goy, Carmen Carrasco Luján, Maddalena Zaglio, Myriam Perregaux, Juliette Janes, Benoît Sagot, Rachel Bawden, Rasul Dent, Oriane Nédey, and Alix Chagué. 2024. *Reconnaissance des écritures dans les imprimés*. In *Humanistica 2024*, OCR, Meknès, Morocco. Association francophone des humanités numériques.
- Thomas-Simon Guelette. 1740. *Première Parade*. Paul Fièvre.
- Walter Haas. 2015. « Déguisé en Suisse » : les « Suisses » de Molière et leur langage. *Littératures classiques*, 87(2):191–189.
- Marie-Christine Hazaël-Massieux. 2008. *Textes anciens en créole français de la Caraïbe: Histoire et analyse*. Editions Publibook.
- Marie-Christine Hazaël-Massieux. 2013a. *Creolica: Revue du Groupe Européen de Recherches en Langues Créoles*.
- Marie-Christine Hazaël-Massieux. 2013b. *Groupe Européen de Recherches en Langues Créoles*.
- Louis Héry. 1883. *Fables Créoles et Explorations Dans l’intérieur de l’île Bourbon: Esquisses Africaines*. J. Rigal.
- Alexander Hull. 1979. On the origin and chronology of the French-based creoles. *Readings in creole studies*, 2:201.
- Alexander Hull. 1993. The transmission of Creole languages. *Atlantic Meets Pacific—A Global View Of Pidginization and Creolization (Selected Papers from the Society for Pidgins and Creole Linguistics)*, pages 391–397.
- Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. *Automatic Language Identification in Texts: A Survey*. *Journal of Artificial Intelligence Research*, 65.
- Charles Jobey. 1860. *L’amour d’un Nègre*. Michel Lévy frères.
- Amr Keleg and Walid Magdy. 2023. *Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification*. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.
- Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. *eScriptorium: An Open Source Platform for Historical Document Analysis*. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.
- Alain Kihm. 2018. *Língua de Preto, the language of the African slave community in Portugal (16th–19th centuries)*. *Language Ecology*, 2(1-2):77–90.
- Ruth King, Terry Nadasdi, and Gary R. Butler. 2004. *First-person plural in Prince Edward Island Acadian French: The fate of the vernacular variant je... ons*. *Language Variation and Change*, 16(3):237–255.
- Thomas Klingler. 2003. *If I could turn my tongue like that: the Creole Language of Pointe Coupee Parish, Louisiana*. LSU Press.
- Jacques Landrecies. 2001. *Une configuration inédite : la triangulaire français-flamand-picard à Roubaix au début du XXe siècle*. *Langage et société*, 97(3):27–69.
- Elizabeth Eva Leach. 2020. *Ripping Romance to Ribbons: The French of a German Knight in The Tournament at Chauvency*. *Medium Ævum*, 89(2):327–349.
- Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022a. *Ancestor-to-Creole Transfer is Not a Walk in the Park*.
- Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022b. *What a Creole Wants, What a Creole Needs*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.
- John M. Lipski. 2001. *Bozal Spanish: Restructuring or creolization? Degrees of Restructuring in Creole Languages*, 22:437.

- Johann-Mattis List. 2024. [Open Problems in Computational Historical Linguistics](#). *Open Research Europe*, 3:201.
- Nikola Ljubesic, Nives Mikelic, and Damir Boras. 2007. [Language Identification: How to Distinguish Similar Languages?](#) In *2007 29th International Conference on Information Technology Interfaces*, pages 541–546.
- Anthony Lodge. 1991. [Molière’s Peasants and the Norms of Spoken French](#). *Neuphilologische Mitteilungen*, 92(4):485–499.
- Anthony Lodge. 2003. *French: From Dialect to Standard*. Routledge, London.
- Leena Löfstedt. 2014. [Notes on the beginnings of Law French](#). *Romance Philology*, 68(2):285–337.
- Pierre Carlet de Chamblain de Marivaux. 1727. *L’île de La Raison*. Paul Fièvre.
- Pierre Carlet de Chamblain de Marivaux. 1735. *La Mère Confidente*. Paul Fièvre.
- Oliver Mayeux. 2019. [Rethinking decreolization: Language contact and change in Louisiana Creole](#).
- John H. McWhorter. 1998. [Identifying the Creole Prototype: Vindicating a Typological Class](#). *Language*, 74(4):788–818.
- John H. McWhorter. 2018. *The Creole Debate*. Cambridge University Press.
- Alfred Mercier. 1881. *L’habitation Saint-Ybars: Ou, Maitres et Esclaves En Louisiane, Recit Social*. Nouvelle-Orléans. Imprimerie franco-américaine (E. Antoine).
- Ingrid Neumann-Holzschuh. 1987. Textes anciens en creole louisianais: Avec introd., notes, remarques sur la langue et glossaire.
- Mikael Parkvall. 1995. The role of St. Kitts in a new scenario of French Creole genesis. *From Contact to Creole and Beyond*.
- Jean-François Regnard. 1690. *Les filles errantes*. Paul Fièvre.
- Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, and Naome Etori. 2024. [Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110.
- Nathaniel Romney Robinson, Matthew Dean Stutzman, Stephen D. Richardson, and David R. Mortensen. 2023. [African Substrates Rather Than European Lexifiers to Augment African-diaspora Creole Translation](#). In *4th Workshop on African Natural Language Processing*.
- Pablo Ruiz Fabo, Delphine Bernhard, and Carole Werner. 2020. [Création d’un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines](#). In *2èmes Journées Scientifiques Du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, pages 34–43, Montrouge, France. CNRS.
- Patrick Sauzet and Guylaine Brun-Trigaud. 2015. [La Lucette de Monsieur de Pourceaugnac : “ Feinte gasconne ”, vrai occitan](#). *Littératures classiques*.
- Charles-Timoléon de Beauxoncles de Sigogne. 1620. Testament d’un Escossois. In *Le Cabinet Satyrique...* Billaine.
- David Baird Smith. 1920. [Le Testament du Gentil Cossoys](#). *The Scottish Historical Review*, 17(67):190–198.
- Anand Sycia. 2017. *French Creoles: A Comprehensive and Comparative Grammar*. Routledge.
- TEI Consortium eds. 2023. [TEI P5: Guidelines for Electronic Text Encoding and Interchange](#).
- Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press.
- Megan Vaughan. 2005. *Creating the Creole Island: Slavery in Eighteenth-Century Mauritius*. Duke University Press, Durham, NC.
- Jules Watteuw. 1927. Pitche.
- Henri Wittmann. 1995. Grammaire comparée des variétés coloniales du français populaire de Paris du 17e siècle et origines du français québécois. *Le français des Amériques, dir. Robert Fournier & Henri Wittmann*, pages 281–334.
- Jonathan Wylie. 1995. [The Origins of Lesser Antillean French Creole: Some Literary and Lexical Evidence](#). *Journal of Pidgin and Creole Languages*, 10(1):77–126.

Vector Poetics: Parallel Couplet Detection in Classical Chinese Poetry

Maciej Kurzynski

Advanced Institute for Global Chinese Studies
Lingnan University, Hong Kong
maciej.kurzynski@ln.edu.hk

Xiaotong Xu

Lingnan University, Hong Kong
xiaotongxu@ln.hk

Yu Feng

Lingnan University, Hong Kong
yufeng2@ln.hk

Abstract

This paper explores computational approaches for detecting parallelism in classical Chinese poetry, a rhetorical device where two verses mirror each other in syntax, meaning, tone, and rhythm. We experiment with five classification methods: (1) verb position matching, (2) integrated semantic, syntactic, and word-segmentation analysis, (3) difference-based character embeddings, (4) structured examples (inner/outer couplets), and (5) GPT-guided classification. We use a manually annotated dataset, containing 6,125 pentasyllabic couplets, to evaluate performance. The results indicate that parallelism detection poses a significant challenge even for powerful LLMs such as GPT-4o, with the highest F1 score below 0.72. Nevertheless, each method contributes valuable insights into the art of parallelism in Chinese poetry, suggesting a new understanding of parallelism as a verbal expression of principal components in a culturally defined vector space.

1 Introduction

Parallelism is a rhetorical and structural device where two poetic verses correspond with each other in terms of syntax, meaning, tone, and rhythm. The most widely recognized form of parallelism in Chinese culture are the red couplets, known as *fai chun* or *chun lian*, traditionally placed on either side of the doors during Lunar New Year, but the most exquisite ones can be found in classical poetry. Parallelism requires careful selection of words to maintain symmetry while expressing complementary or contrasting ideas, making it a hallmark of linguistic artistry in Chinese poetic craft.

Consider the following, classic example of a pentasyllabic parallel couplet:

山光悦鸟性，
潭影空人心。

*The mountain light pleases the bird's nature;
The pond's reflection empties the human heart.*

Each character (syllable) in the first line corresponds syntactically and semantically with a character at the same position in the second line: 山 (“mountain”) and 潭 (“pond”) both describe natural elements, while 光 (“light”) and 影 (“reflection”) refer to visual phenomena. Similarly, 悦 (“pleases”) and 空 (“empties”) are verbs that describe the effects of these phenomena on 鸟性 (“bird’s nature”) and 人心 (“human heart”), respectively. This correspondence creates a balanced and harmonious structure, parallel in both form and meaning.

While not all Chinese poetry contains parallel structures, one of the most rewarding aspects of reading regulated verses like the one shown above lies in uncovering the intricate correspondences between characters, words, and entire lines. These may be readily apparent, as in our example, or more subtly concealed, drawing upon obscure references and tropes from classical literature that only the most knowledgeable readers will fully appreciate.

In this paper, we report on our attempts to develop a computational method for detecting parallelisms in poetic verses. This complex task involves numerous linguistic dimensions, such as word segmentation, semantic comparison, and part-of-speech alignment, stretching to the limit the abilities of language models and their emergent properties. Yet beyond the technical challenges, computational studies of parallelism can also shed new light on the broader questions of how meaning is generated through language and how humans perceive, structure, and understand the world through patterns and relationships.

2 Previous Work

Modern scholars and traditional *literati* alike have long debated what constitutes a parallel couplet, identifying various categories of parallelisms. In the chapter “Parallel Phrasing” from his *Literary*

Mind and the Carving of Dragons, Liu Xie (460-522) distinguishes four kinds of couplets: verbal (言对), material (事对), antithetical (反对), and direct (正对) (Liu, 2015). A Japanese scholar monk, Kūkai (774-835), proficient in Chinese classical literary criticism, identifies as many as 29 different types of couplets (Kūkai, 1975). Wang Li, a prominent modern Chinese linguist, argues that the classification of words—pairing nouns with nouns, verbs with verbs, etc.—is foundational to understanding parallelism (Wang, 1979). Andrew Plaks views parallelism not merely as a stylistic tool but as a fundamental mode of textual organization and argumentation (Plaks, 1990). Zhang Longxi offers a comparative East-West view on parallelism as a way of reasoning common to all humans as embodied beings (Zhang, 2021). Cai Zong-qi traces the development of parallel patterning in Chinese literature beginning with pre-Qin writings and becoming prominent during the Six Dynasties (Cai, 2022).

From the natural language processing (NLP) perspective, research on parallelism in Chinese poetry has drawn relatively little attention. Lee et al. have applied POS-tagging to detect syntactic parallelism in Tang poetry, confirming traditional views that the middle couplets (2nd and 3rd) in a 4-couplet poem are more often syntactically parallel than the outer couplets (1st and 4th) (Lee et al., 2018). Few other studies focus on the automatic generation of parallel couplets (Yuan et al., 2019; Song, 2022; Qu et al., 2022). Projects like the *Chinese Classical Poetry Matching Dataset (CCPM)* provide structured data that facilitates the study of semantic relationships in poetry, enabling models to better understand the nuances of parallelism by matching ancient Chinese lines with their modern translations (Li et al., 2021). Another prominent approach is the use of transformer-based models, such as the *TransCouplet* model, which leverages a fusion of embeddings, including glyph, pinyin, and part-of-speech embeddings, to capture the syntactic and phonetic nuances of Chinese couplets. The model employs both a transformer encoder and decoder to generate the second line of a couplet given the first, ensuring that the generated line adheres to the linguistic rules of parallelism (Chiang et al., 2021). Yet another innovative method is the *CoupGAN*, which utilizes a generative adversarial network (GAN) for couplet generation. This model focuses on learning the intricate balance between semantic meaning and structural symmetry

that characterizes high-quality couplets (Qu et al., 2022). All of these studies contribute to the ongoing efforts to apply advanced machine learning techniques to the generation and analysis of Chinese parallel couplets, pushing the boundaries of how well machines can replicate this sophisticated form of poetic expression.

3 Dataset

Our manually-annotated test dataset, used to compare different classification methods, contains 6,125 pentasyllabic couplets divided into two classes: parallel (2,139 samples) and non-parallel (3,986 samples). All couplets come from the so-called “Six Dynasties” (222-589), a transformative period in the history of Chinese literature which witnessed the development of parallel poetry, ultimately leading to the elegant regulated verses of the Tang (618-907) and the following eras. The manual annotation ensures a good amount of reliable data to conduct evaluation. Since each method has been using a different training dataset, we provide detailed descriptions of the training process in the relevant subsections.

4 Experiments

4.1 Baseline: Verb Matching

Our baseline method focused on the syntactic positioning of verbs. For training, we have used a manually-annotated dataset of pentasyllabic poetry containing 7,733 verses, with each verse labeled with binary code indicating a verb (1) or non-verb (0) at each position. For example, the line 翩翩吹我衣 (“blowing fluttringly on my clothes”) would be labeled “00100,” with “1” indicating the verb 吹 in the third position. The *SikuBERT* model¹ (109M parameters), pre-trained on a large collection of pre-modern Chinese texts, was then fine-tuned for one epoch on this labeled dataset. We used Adam optimizer, batch size of 16, and learning rate 2e-5. During inference, if verbs were found at corresponding positions in both lines, we considered them as parallel. For example, in the couplet 人来水鸟没, 楫度岸花沈 (“as people arrive, water birds vanish; as oars pass, riverside flowers submerge”), the model correctly identified the verbs that appear at the second and fifth positions in both lines (“01001”).

¹<https://huggingface.co/SikuBERT/sikubert>

4.2 Integrated Semantic and Syntactic Analysis

The second method took a more holistic approach, integrating three distinct models: word segmentation, part-of-speech tagging, and character-to-character semantic matching.

The **word segmentation** model split poetic lines into words, which is essential for Chinese, where word boundaries are not explicitly marked, and where a word can consist of one or more characters. We fine-tuned *SikuBERT* on a dataset of pentasyllabic poetry containing 540k couplets, which had been pre-processed by segmenting each couplet into individual words. To do that, we have built a list of words from the *Hanyu Da Cidian*, or the *Grand Chinese Dictionary*. For example, the couplet 猛虎凭林啸, 玄猿临岸欢 (“Fierce tigers roar from within the forest, mysterious monkeys rejoice by the riverbank”) would be segmented into (first line) 猛虎 “fierce tigers” | 凭 “leaning” | 林 “the forest” | 啸 “roar” and (second line) 玄猿 “mysterious monkeys” | 临 “approaching” | 岸 “the riverbank” | 欢 “rejoice.” We labeled each character as 1 (beginning of the word), 0 (middle or end of the word) or 2 (*BERT* special tokens [CLS] and [SEP]). In the above example, the two lines align perfectly, segmented into 2+1+1+1 structures (or “2101112101112” using our notation). We fine-tuned *SikuBERT* on thus pre-processed couplets (including both parallel and non-parallel ones) for one epoch, with batch size of 16, Adam optimizer, and the learning rate $2e-5$. In inference, we segmented each test couplet and measured the segmentation alignment between the two lines; since the first character is always the beginning of a word, each correct matching after the first character would increase the segmentation score by 0.25.

The **part-of-speech** (POS) tagging model attached tags to each character, providing structural information about the sentence. For example, in the couplet 挥汗辞中宇, 登城临清池 (“Wiping sweat as he leaves the central palace, ascending the ramparts to gaze upon the clear pond”), both lines match perfectly as V+N+V+ADJ+N. For this task, we used a third-party *RoBERTa* model pre-trained on Classical Chinese texts for POS-tagging and dependency-parsing.² Each correct pair-wise matching increased the POS score of a couplet by 0.2.

The **semantic matching** model, finally, com-

pared characters between the two lines to assess semantic relationships. For instance, in the couplet 共矜红颜日, 俱忘白发年 (“Together we cherished the days of youth, equally we forgot the years of grey hair”), all characters in the first line match semantically with those in the second line: adverbs with adverbs, colors with colors, time markers with time markers, etc. Conversely, in 常恐寒节至, 凝气结为霜 (“Always fearing the arrival of cold seasons; the air thickens and turns to frost”), there are no semantically matching characters.

The semantic-matching model required more work than the other two. First, we used *SikuBERT* to generate embeddings (last hidden states, which contain the deep semantic representations) of each character in a couplet. A smaller model, placed on top of the *SikuBERT* encoder, then constructed a rich feature set by concatenating the embeddings of the two corresponding characters as well as their difference. This combined vector (of dimensionality $768 * 3 = 2304$) was passed through a classification layer that predicted whether the character pair matched semantically. By repeating this process across all corresponding character pairs in the couplet, the model generated a final score (the average of pairwise scores) that determined the overall semantic parallelism. The positive training examples were derived from the parallel couplets (the second and third line pairs, or the middle couplets) gathered from the regulated pentasyllabic verses composed in the Tang and the following dynasties, including 270k examples. An equal number of negative examples was generated synthetically by randomly redistributing from 2 to 5 characters within either line of the parallel couplet. For instance, the positive example 青溪通小市, 香稻满平畴 (“A blue stream leads to a small market; Fragrant rice fills the flat fields”), where all characters match semantically (“11111”), would be transformed into a negative example 通青溪小市, 香稻满平畴 (lit. “Leads blue stream small market; Fragrant rice fills the flat fields”), labeled as “00011.” In this case, we shuffled the first three characters in the first line.

Having trained the three models, in inference the scores were then averaged, and if the final score exceeded a predefined threshold (in our case, the best results were achieved with threshold of 75%), the couplet was classified as parallel.

4.3 Difference-Based Character Embeddings

This method focused on the idea of character-to-character differences in embedding space, us-

²KoichiYasuoka/roberta-classical-chinese-large-upos

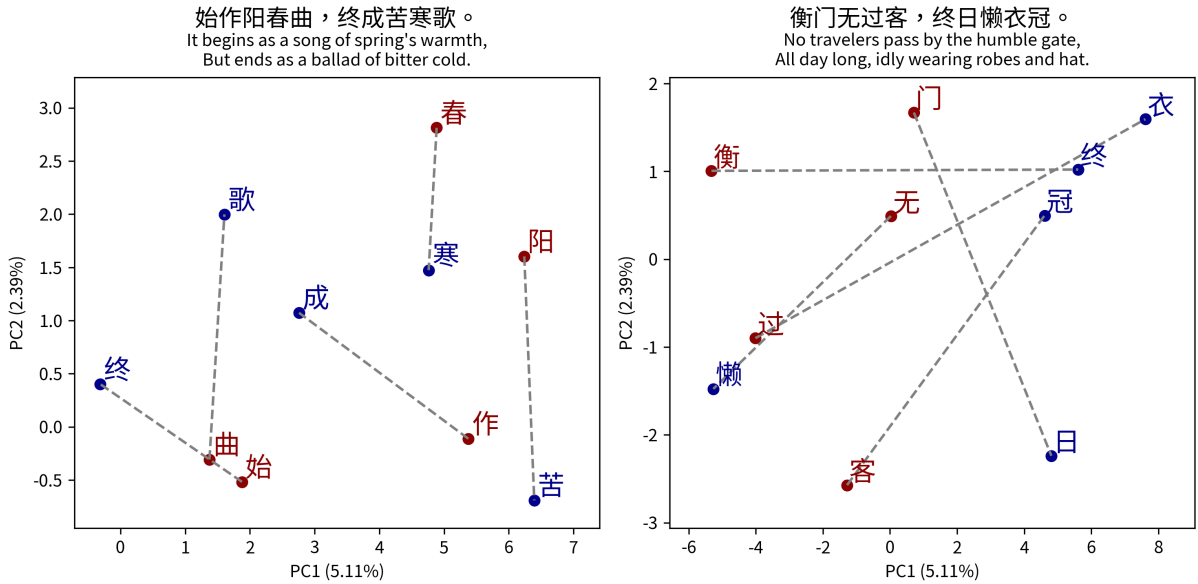


Figure 1: 2-dimensional PCA projections of character embeddings of parallel (left) and non-parallel (right) couplets. Embeddings (last hidden states, $dim = 768$) were generated by *SikuBERT*. We randomly selected 600 couplets to find principal components. Although the first two PCs explain relatively little variance, some parallel relationships can be recognized. For example, in the parallel couplet (left), 阳 (“warmth”) is to 苦 (“bitter”) as 春 (“spring”) is to 寒 (“cold”), and 作 (“to make”) is to 成 (“to become”) as 始 (“to begin”) is to 终 (“to end”).

ing them to distinguish between parallel and non-parallel couplets. For each training couplet, contextual embeddings were generated for all characters with the help of *SikuBERT*. The differences between the contextual embeddings of corresponding characters from the two lines were then calculated and stored in a list as “parallel differences.”

To introduce contrast, the characters in one of the lines were shifted by one or more positions, and the same embedding difference calculation was performed. These results were labeled as “non-parallel differences,” as shifting disrupts the proper alignment necessary for parallelism while preserving semantic connectedness between the lines. A one-layer classifier with RELU non-linearity was then trained for one epoch using both parallel and non-parallel differences (270k examples each).

Once trained, the classifier was applied to our test couplets. If the couplet consisted only of “parallel differences,” we labeled the couplet as parallel. This method aimed to capture the essence of parallelism by focusing on the relative positions and differences between characters in the embedding space (Figure 1).

4.4 Structured Examples

The next method leveraged the structured nature of regulated poetry from the Tang Dynasty onwards,

particularly the *wulü* (five-character eight-line regulated verse). As mentioned before, these poetic forms prescribe parallel lines at specific positions, typically in the second and third couplets. To harness this structural consistency for classification, we collected couplets from publicly available datasets of classical Chinese poetry. Positive examples (270k) were taken from the second and third couplets, while the first and fourth couplets served as negative examples (270k).³ We fine-tuned *SikuBERT* on this dataset for one epoch, with a batch size of 16, learning rate of $2e-5$, and Adam optimizer.

4.5 GPT-Guided Evaluation

The final method leveraged the capabilities of state-of-the-art AI in the form of GPT-4o model (gpt-4o-2024-05-13) accessed over OpenAI API. GPT-4o is a large, multilingual, and multimodal generative transformer model which can be requested to perform various tasks through prompting. This method involved combining a list of calibration examples and a target couplet within a prompt:

Few-Shot Learning: The GPT-4o model was first provided with examples of couplets: in one version, we provided only two examples, one par-

³Some of the first and fourth couplets *are* parallel, a problem explored in the Discussion section below.

allel and one non-parallel, in another as many as ten examples (five parallel and five non-parallel). All examples included a simple analysis and the final judgment based on this analysis. This part of the prompt calibrated the model’s understanding of parallelism in classical Chinese poetry, particularly in terms of grammatical, syntactic, and semantic matching. We provide the 2-shot prompt (original and translation) in the appendix.

Couplet Evaluation: The second part of the prompt included the target couplet. For each couplet, the model was supposed to follow the instructions from the first part of the prompt by providing the couplet’s structural and semantic analysis and then deciding whether it exhibited parallelism. For example, given the lines 驰波催永夜，零露逼短晨 (“Rushing waves hasten the endless night, Scattered dew pressures the brief morning”), the model provided the following explanation (translated here into English, original Chinese below):

“Rushing waves” and “scattered dew” match: both are natural scenes; “hasten” and “pressure” match: both are verbs that imply driving or prompting; “endless night” and “brief morning” also match: one represents a long night and the other a short morning, creating a contrast. Words at each corresponding position in the two lines match semantically, syntactically, and grammatically, thus forming a neatly balanced couplet.

Decision: parallel.

「驰波」与「零露」匹配：都是自然景象。「催」和「逼」匹配：都是动词，表示推动、促使的意思。「永夜」和「短晨」也匹配：一个表示长夜，一个表示短晨，形成对比。两句在每个位置上的词语在语义、句法和语法上都相对应，所以构成工整对仗的对联。

判断：是对联。

5 Results

Despite expectations, the GPT-guided evaluation did not significantly outperform more traditional methods (Table 1). While GPT-4o boasts extensive training data, reinforcement learning from human feedback (RLHF), and a vast parameter count, its performance was comparable to simpler, structured methods. We encountered a number of confusing cases where the final classification seemingly contradicted the preceding analysis, or where the

analysis was hallucinated. As the model is non-deterministic, in some cases GPT-4o generated two contradictory interpretations. The most common errors occurred when the model did not know how to parse a poetic line. Consider the following example of a hallucinated reasoning generated in response to the couplet 淳于前行酒，雍门坐相和 (“Chunyu presents wine as he walks ahead, Yongmen sits and joins in harmony”):

Analysis: “Chunyu” and “Yongmen” are both names of people, matching each other. “Walking ahead” [前行] and “sitting” [坐] express different actions, one walking and the other sitting, forming a neat antithetical movement. “Wine” [酒] and “harmonizing” [相和] represent two scenes related to drinking and music, basically corresponding to each other.

「淳于」和「雍门」都是人名，匹配。「前行」和「坐」表达了不同的动作，一个是行走，一个是坐着，动作对仗工整。「酒」和「相和」表示的是与饮酒和音乐相关的两个场景，基本对应。

While the model correctly identified the proper names, it considered the two verses as parallel despite the syntactic mismatch in the last three characters, comparing words of different lengths. We believe that further prompt engineering could mitigate such errors and improve performance.

The verb-based method, on the other hand, struggled with lines where verbs were either absent or not central to the couplet’s structure, leading to misclassifications. Consider the couplet 春草郁青青，桑柘何奕奕 (“spring grass flourishes, how lush and vibrant; mulberry and elm trees, how splendid and bright”), which features no verbs (“00000”) but still forms a parallel couplet. Moreover, even when verbs were present at identical positions, semantic mismatches could occur, such as in 边地无高木，萧萧多白杨 (“There are no tall trees in the borderland; there are many rustling poplars”), where the verbs align, but the initial disyllabic words (边地 “borderland” and 萧萧 “rustling”) do not match semantically.

The integrated approach remained heavily dependent on the availability and quality of training data, and the compounded errors from individual models led to numerous misclassifications. For example, the word-segmentation model alone could not account for semantic mismatches, as seen in 误落尘网中，一去三十年 (“Mistakenly caught in the

Table 1: Results of Classification Methods. The methods include Verb Matching (VM), which focuses on the syntactic positioning of verbs; Semantic-Syntactic Integration (ISSA), which combines semantic matching with syntactic analysis (POS + word segmentation); Embedding Differences (ED), which utilizes character embedding differences; Structured Examples (SE), which distinguishes between 2nd and 3rd (mostly parallel) couplets and 1st and 4th (mostly non-parallel) couplets from regulated poetry; and GPT Evaluation (GPT-4o), which leverages a large conversational transformer model.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
VM (baseline)	76.5	72.3	55.0	62.5
ISSA	72.5	68.2	48.0	56.3
ED	75.9	72.0	53.6	61.4
SE	79.1	69.9	72.7	71.3
GPT-4o (2-shot)	79.0	70.7	68.4	69.5
GPT-4o (10-shot)	79.6	73.7	64.4	68.8

web of the mundane, Gone for thirty years”), where the word segmentation is consistent (1+1+2+1), but semantics differ. Likewise, the part-of-speech model would (wrongly) reduce the couplet score in cases where non-matching individual characters formed larger matching units: in the couplet 君居北海阳，妾在江南阴 (“You reside in the north of the North Sea; I dwell in the south of the Yangtze”), the POS tags of individual characters differ in 北海 (North Sea, labeled as adjective+noun) and 江南 (Yangtze’s South, labeled as noun+adjective), even though the two words match semantically.

Finally, although the structured examples approach should have benefited from the fact that the regulated poetry from the Tang onwards contains parallel lines at prescribed positions, we have found many exceptions in the existing datasets, in particular verses from less known poets that do not conform with the stringent formal rules, such as the inner couplet 隔邻呼酒伴，对坐且开尊 (“Calling a drinking companion from next door; They sit facing each other, opening a jar of wine”). Similar examples might have confused the model during training, leading to lower score. It is worth noting that despite such shortcomings, this approach achieved higher F1 score than GPT-4o, and further filtering of the training data might improve the results. On the other hand, it offers less explainability than GPT.

6 Discussion

6.1 Parallelism and Subjectivity

In our evaluation of parallel couplets, it became evident that the notion of parallelism is not always binary—many couplets do not simply fall into “parallel” or “non-parallel” categories. Instead, paral-

lelism often manifests on a spectrum. We observed instances of “almost parallel” couplets where, despite imperfect POS matching, the semantic correspondence between two lines suggested itself: 昔闻汾水游，今见尘外骖 (“Once I heard of travels by the Fen River, Now I see a horse beyond the dusty realms”).

Likewise, the subjective nature of interpreting poetic texts posed a significant challenge in our study. A creative reader can find parallel relationships between any two randomly chosen verses. We frequently encountered disagreements among annotators regarding the classification of certain couplets. These disagreements highlight the inherent interpretive flexibility in literary analysis, where different readers may perceive varying degrees of parallelism based on their linguistic intuition, cultural background, and personal experience.

6.2 Vector Poetics

The concept of “vector poetics” represents a promising avenue explored in this project, drawing inspiration from foundational work in natural language processing. Seminal studies of word embeddings, such as word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) papers, demonstrate how parallel relationships between words in a given dataset can be quantitatively modeled (“queen” is to “king” as “woman” is to “man,” e.g.). We suggest that this analogy can be usefully extended to sets of words—or entire poetic lines—in parallel couplets (Figure 1). Consider the transformation:

$$\text{春} - \text{生} + \text{死} \approx \text{秋} \quad (1)$$

Translating as “Spring – Life + Death \approx Autumn,”

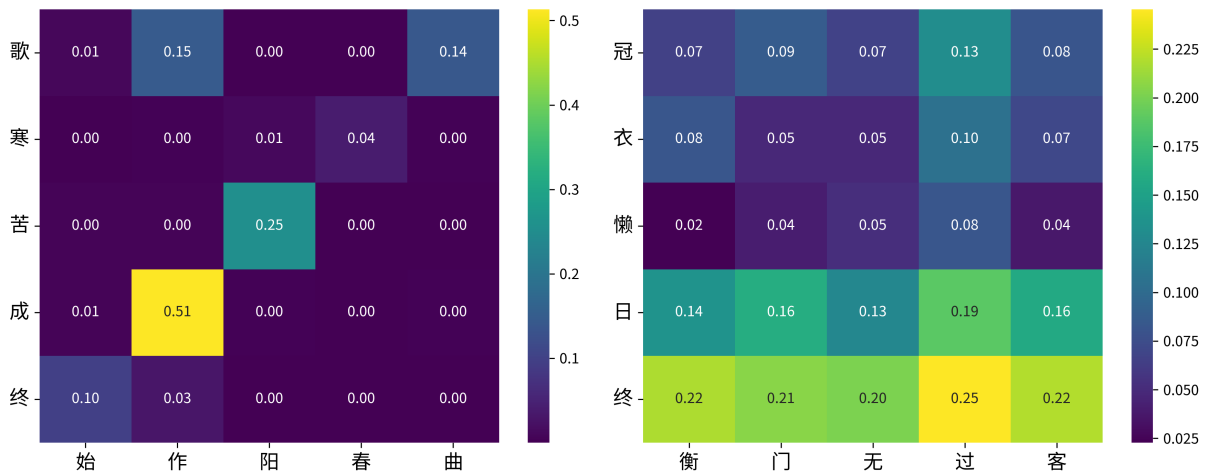


Figure 2: Attention heatmaps comparing parallel and nonparallel couplets (examples identical with Figure 1). Left: Heatmap showing focused attention along the diagonal in a parallel couplet, indicative of strong structural and semantic alignment learned by the model from parallel examples. Right: Heatmap for a non-parallel couplet with diffused attention, highlighting the lack of structured alignment between lines. Attention scores have been averaged (both directions) over all heads from the top layer of *SikuBERT*, which has been fine-tuned on parallel and non-parallel couplets (Structured Examples approach).

the equation conceptualizes the cyclical nature discovered by poets in both seasons and life. Importantly, such equivalences will hold only in cultural contexts where “Spring” is to “Life” as “Autumn” is to “Death,” that is, where these concepts align along the same “principal cultural components.” Vector poetics would thus re-interpret poetic parallelisms as manifestations of cultural dimensions in which two things are (subjectively) perceived as either identical or opposite. This also entails a quantitative re-interpretation of poetic creativity: a search within a historically and culturally defined vector space, and a reshaping of that space through discovery of new similarities and differences between things in the world.

Finally, our analysis points to the innovative use of attention mechanisms within transformer-based models, which can visualize the interaction between poetic lines (Figure 2). Heatmaps generated from the attention scores in a fine-tuned *SikuBERT* reveal distinct patterns: in parallel couplets, there is a noticeable concentration of attention along the diagonal, indicating a structured and direct interaction between corresponding characters. Conversely, in non-parallel couplets, attention is more diffusely spread across the matrix, indicating a weaker alignment and less structured interaction between the lines. Notice that these specific attention patterns have been inferred by the model during training. While more work needs to be done in this direction, currently available results seem to offer further evi-

dence that parallelism is not only a formal device at work within a text but also a cognitive device that guides the reader’s interpretive engagement.

Limitations

Despite the advancements detailed in this study, several limitations remain. Traditional Chinese poetry, particularly from the classical period, adheres to strict patterns of alternating tones, known as *pingze* (平仄), with *ping* (平) tones being level and *ze* (仄) tones oblique. These tonal patterns, crucial for the aesthetic and rhythmic quality of Chinese poetry, add yet another layer of complexity to parallelism detection but are not accounted for in our current computational models, which generally lack the capability to analyze tonal features (Liu et al., 2018). Additionally, our approach overlooks the ancient pronunciation of Chinese characters, which is vital for understanding the original phonetic environment and rhyming schemes of the texts (Tharsen, 2022). This oversight may lead to potential misinterpretations or missed phonetic parallelisms that would have been evident to contemporary audiences. Finally, the pre-trained *SikuBERT* model used in our study is relatively small by contemporary standards; larger models are likely to yield better performance. We are going to address these shortcomings in our future studies.

Acknowledgments

The authors would like to thank the NLP4DH 2024 conference organizers and the two anonymous reviewers for their useful comments and feedback.

References

- Zongqi Cai. 2022. Six dynasties parallel prose: Descriptive and expository. In Zong qi Cai, editor, *How to Read Chinese Prose*, pages 210–225. Columbia University Press, New York.
- Kuan-Yu Chiang, Shihao Lin, Joe Chen, Qian Yin, and Qizhen Jin. 2021. [Transcouplet: Transformer based chinese couplet generation](#). *arXiv preprint arXiv:2112.01707*.
- Kūkai. 1975. *Bunkyo Hifuron*. People’s Literature Publishing House.
- John Lee, Yin Hei Kong, and Mengqi Luo. 2018. [Syntactic patterns in classical chinese poems: A quantitative study](#). *Digital Scholarship in the Humanities*, 33(1):82–95.
- Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. [CCPM: A Chinese Classical Poetry Matching Dataset](#). *arXiv preprint arXiv:2106.01979*. Accessed: 2024-08-19.
- Chao-Lin Liu, Thomas J Mazanec, and Jeffrey R Tharsen. 2018. Exploring chinese poetry with digital assistance: Examples from linguistic, literary, and historical viewpoints. *Journal of Chinese Literature and Culture*, 5(2):276–321.
- Xie Liu. 2015. *The Literary Mind and the Carving of Dragons*. The Chinese University Press.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.
- Andrew H. Plaks. 1990. [Where the lines meet: Parallelism in chinese and western literatures](#). *Poetics Today*, 11(1):105–127.
- Qian Qu, Jiancheng Lv, Dayiheng Liu, and Kexin Yang. 2022. [Coupgan: Chinese couplet generation via encoder–decoder model and adversarial training under global control](#). *Soft Comput.*, 26(15):7423–7433.
- Yan Song. 2022. [Chinese couplet generation with syntactic information](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6436–6446, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Jeffrey R. Tharsen. 2022. [From form to sound: visual and aural representations of premodern chinese phonology and phonorhetoric with applications for phonetic scripts](#). *International Journal of Digital Humanities*, 4(1-3):115–129.
- Li Wang. 1979. *A Brief Introduction to the Metric Patterns of Chinese Poetry*. Beijing Publishing House.
- Shengqiong Yuan, Luo Zhong, Lin Li, and Rui Zhang. 2019. [Automatic generation of chinese couplets with attention based encoder-decoder model](#). *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 65–70.
- Longxi Zhang. 2021. Parallelism and antithesis: Structural principles in the mind and in literature from a chinese perspective. *European Review*, 29(2):274–284.

A Appendix (2-shot GPT-4o prompt)

[system] You are a professor of Chinese literature and an expert in ancient poetics.

[user] Please determine if the following two five-character lines form a parallel couplet. The requirement for a parallel couplet is that each word in the first line corresponds syntactically, grammatically, and semantically with a word in the second line (there are many methods of correspondence: the two words can respectively represent time and space, morning and evening, life and death, youth and harsh winter, static and dynamic, two different places, place nouns and time nouns, etc.). You need to provide a concise analysis, then make a decision based on this analysis. The decision can only be ‘is parallel’ or ‘is not parallel.’

[user] Couplet: Observing things arouses sorrow, looking after oneself recognizes the source of worries.

[assistant] Analysis: “Observing things” corresponds to “looking after oneself”: one is about observing the outside world, the other is about caring for one’s inner world (both are one-character verb plus one-character noun structures), “arouses” and “recognizes” are both verbs, “sorrow” refers to a sad mood, “source of worries” refers to a sorrowful mood, both match. Decision: Is parallel.

[user] Couplet: Sitting quietly all around, I listen to the song above my hall.

[assistant] Analysis: “Four” and “listen” do not match (one is a number, the other is a verb), “sit” and “I” do not match (one is a verb, the other is a pronoun), “and” and “hall” do not match (one is a

function word, the other is a noun). Decision: Is not parallel.

[system] 你是中国文学的教授和古代诗学的专家。

[user] 请判断下面的两个五言句是不是工整对仗的对联。对联的要求是前句和后句的每一个词语在句法、语法和意义上都相对应（对应的方法有很多：两个词可以分别表达时间与空间、早上与晚上、生与死、春天与冬天、静态与动态、两个不同的地方、地点名词与时间名词等等）。请提供简洁的分析，然后根据这个分析进行判断。判断只能是「是对联」或者「非对联」。

[user] 句对：览物起悲绪，顾己识忧端。

[assistant] 分析：「览物」对应着「顾己」：一个是指观赏外面的世界，一个是指照顾内心的世界（两个都是一个字的动词加一个字的名词），「起」与「识」都是动词，「悲绪」是指悲伤的心情，「忧端」是指忧伤的心情，都匹配。判断：是对联。

[user] 句对：四坐且莫喧，听我堂上歌。

[assistant] 分析：「四」与「听」不匹配（一个是数字，一个是动词），「坐」与「我」不匹配（一个是动词，一个是代名词），「且」和「堂」不匹配（一个是功能词，一个是名词）。判断：非对联。

Adapting Measures of Literality for Use with Historical Language Data

Adam Roussel

Department of Linguistics
Ruhr University Bochum
roussel@linguistics.rub.de

Abstract

This paper concerns the adaptation of two existing computational measures relating to the estimation of the literality of expressions to enable their use in scenarios where data is scarce, as is usually the case with historical language data. Being able to determine an expression's literality via statistical means could support a range of linguistic annotation tasks, such as those relating to metaphor, metonymy, and idiomatic expressions, however making this judgment is especially difficult for modern annotators of historical and ancient texts. Therefore we re-implement these measures using smaller corpora and count-based vectors more suited to these amounts of training data. The adapted measures are evaluated against an existing data set of particle verbs annotated with degrees of literality. The results were inconclusive, yielding low correlations between 0.05 and 0.10 (Spearman's ρ). Further work is needed to determine which measures and types of data correspond to which aspects of literality.

1 Introduction

Though it is usually taken as a given that an expression's 'literal' meaning is readily identifiable, the notion is more ambiguous and complex than it initially seems. As Lakoff (1986) explains, there are several different senses that may apply when we describe something as 'literal'. When an expression is used in a literal sense, we might mean that it is used to mean the thing it *usually* means, whether relative to the broader language community or to a narrower domain-specific lect. Alternatively, we might mean that the meaning of the whole expression corresponds to a systematic combination of its constituent parts, i.e. that the expression's meaning is compositional. Or we might be saying something about what an expression is not doing: It's not an instance of irony, metonymy, metaphor, or any other sort of context-dependent extension of its *minimal*

meaning. Yet generally, in linguistic research, we tend to regard these as being equivalent.

It makes a difference which definition you choose and how you regard literality: In Gibbs Jr. et al. (1993) annotators were presented with annotation guidelines highlighting one of Lakoff's definitions of literality and asked to judge the literality of a range of expressions. The result was that some expressions were judged as having significantly different properties, depending on the literality definition in use.

There are a range of semantic annotation tasks which, implicitly or explicitly, require the annotator to make a judgment as to the literality of an expression in a given context: This could be the case in compositionality annotation of multiword expressions (MWEs) or the annotation of idioms or metaphors, which must be distinguished from what they are not: literal usages.

Yet, when annotators are working with historical language data, the ambiguity of these distinct variations of literality, which could otherwise be mitigated with well-written annotation guidelines, is coupled with a lack of linguistic intuition, which is what semantic annotation tasks usually rely on.

In the larger research project in which this paper is situated, it is our aim to develop statistical measures to estimate the degree of literality of particular usages of a given expression in order to support annotators by giving them tools to compensate the lack of linguistic intuitions for historical language varieties. To this end, here we:

- identify existing measures for estimating the degree of literal and non-literal language use,
- adapt the relevant measures for use with historical language data, and
- present the results of a comparison with the original formulation of these measures.

2 Related Work

The theoretical descriptions of literality outlined above highlight multiple aspects which characterize it: The literal meaning is the one that is *conventional* or at least *typical* for that expression, or its the one that is involved when an expression is read compositionally, or its the *minimal* meaning, which an expression is thought to have devoid of any context, its context-free interpretation. Previous approaches have tended to focus either on an expression’s compositionality or its conventionality when measuring the degree to which it is literal or non-literal.

Much of the computational work that deals with the measurement of degrees of compositionality and literality focuses on particular classes of MWEs. Most commonly these are nominal compounds, as in [Schulte im Walde et al. \(2013\)](#), [Weeds et al. \(2017\)](#), and [Cordeiro et al. \(2019\)](#), but others address compositionality in MWEs more broadly, as in [Salehi et al. \(2015\)](#) or [Savary and Cordeiro \(2017\)](#). In general, an expression is taken to be *compositional* to a greater degree when the semantics of the whole are more similar to a systematic combination of the expression’s parts. The composition function is usually an additive model ([Mitchell and Lapata, 2008](#)), which assigns a weight to each vector before adding them together. This approach to compositionality requires a representation for both the expression as a whole as well as for the parts individually, so it tends to work best with a fixed inventory of expressions to be analysed, since you need to know which expressions to combine before training a semantic representation ahead of time. Since our data set doesn’t concern a fixed set of expressions, we plan to address the aspect of compositionality in future work.

There are a range of studies which deal with the annotation and classification of figurative language which are also relevant, such as those concerning metaphor, metonymy, irony, and idioms. In all of these cases, ‘literal’ is defined negatively, as the class not sought, the normal and default assumption. As such, the features that would characterize literality are not modelled directly.

The set of studies that come closest to providing an account of literality itself are those that concern idiomatic expressions. Many of the features discussed in these studies reflect the notion that literal mentions are somehow ‘typical’ of the expression in question and non-literal mentions are thus ‘atyp-

ical’. They differ in how this typicality is modelled. [Sporleder and Li \(2009\)](#) use an unsupervised approach that applies the notion of *lexical cohesion*, operationalizing this as the mean similarity of the terms within a particular window. The candidate expression is then removed and the cohesion is calculated without it. When the level of cohesion increases upon removal of the candidate expression, it suggests that this is an atypical context for that expression, and the instance is classified as idiomatic or non-literal. [Ehren \(2017\)](#) describes an extension of this approach, which replaces the “normalized Google distance” of the original, which has a very low degree of reproducibility, with similarities between word vectors. This approach forms the basis for the cohesion measure we employ below.

[Socolof et al. \(2022\)](#) also address the identification – or rather the *characterization* – of idioms: Their study is more of a characterization because they consider idioms to exist on a spectrum together with novel metaphors, collocations, and ordinary literal language. Thus it isn’t a distinct class as a classification task would suggest. All of these expressions can be related to one another along two orthogonal axes of conventionality and contingency, where conventionality describes the extent to which words are used in their “usual” or “typical” sense, and contingency refers to the tendency for words to be used in a particular, fixed context. The dimension that broadly differentiates literal and non-literal usages is that of conventionality. The conventionality measure that we adapt in this work stems from this study.

3 Methods

3.1 Evaluation data set

In order to evaluate these measures with regard to how well they reflect our intuitions as to the literality of expressions in general, we compare their output to a modern data set of German particle verbs (grammatical constructions consisting of a verb and a separable particle) annotated with literality ratings on a scale from 0 ‘literal’ to 5 ‘non-literal’ ([Köper and Schulte im Walde, 2016](#)). We evaluate against this data set not because of any interest in particle verbs in particular, but because it is the only data set we are aware of that contains scalar ratings of literality rather than a binary classification. The data set consists of German sentences containing particle verbs, as in examples (1)–(3), with roughly 50 sentences for each

one. Since some of the particle verbs were not so frequent in the corpus from which the sentences were extracted, some have fewer than 50 sentences, and we omit those instances with less than 5 sentences, since no reasonable comparison between the instances is possible in that case. Thus we have for this study 155 distinct particle verbs, with 6426 sentences total and 41.5 sentences per lemma on average.

The literality ratings for the examples (1)–(3) are also given below. We include a clearly literal instance (1), a clearly non-literal one (2), and one marginal one (3). Each instance was rated by three raters, with high correlation between them, raters 1 and 2 Pearson’s $\rho = 0.959$, 2 and 3 = 0.943, and 1 and 3 = 0.932, though the corresponding agreement appears moderate with Fleiss’ $\kappa = 0.35$. In the original study, these ratings were combined into two bins, literal and non-literal, with an agreement of Fleiss’ $\kappa = 0.70$ for this classification-oriented setting. In examples (1)–(3) we include the ratings averaged across all three raters.

- (1) Dazu untere Äste kräftig **abklopfen** und herabfallende Läuse auf einem Stück Papier oder Karton auffangen.
‘To that end **pound** heavily on the lower branches and catch the lice that fall down on a piece of paper or cardboard.’
0 \Rightarrow **literal**
- (2) Bin ein alter Bücherwurm und hab meine Spezialadressen **abgeklopft**.
‘I am an old bookworm and have **checked** (lit. knocked on) my special addresses.’
5 \Rightarrow **non-literal**
- (3) Kommerzielle Seiten werden nur in Ausnahmefällen **aufgenommen**.
‘Commercial pages will only be **taken up** in exceptional cases.’
3.67 \Rightarrow **non-literal**

We observe that the ratings are strongly biased towards the extremes of the scale, with 0 and 5 being the most common ratings overall and only very rare instances of 2 or 3, suggesting that the intended use in a classification task was part of the instructions given to annotators.

The data for some lexemes contain very few non-literal instances, and others are more mixed.

3.2 Adapting literality measures

In this study, we adapt two measures: One, [Ehren \(2017\)](#), relates to lexical cohesion, is an embeddings-based version of [Sporleder and Li \(2009\)](#)’s original version, which relied on “normalized Google distances”. The second, conventionality ([Socolof et al., 2022](#)), compares a single instance of an expression with a set of background instances, measuring the degree to which this instance deviates from the general tendency of the background set.

The original formulations rely on resources that are often available, indeed abundant, for modern languages, large corpora derived from collections of unstructured text scraped from the web. The word2vec vectors used in [Ehren \(2017\)](#) tend to require about 1 billion tokens before they are of usable quality ([Sahlgren and Lenci, 2016](#)), and the BERT model used in [Socolof et al. \(2022\)](#) would have been trained on 4 billion words from Wikipedia in addition to other sources.

While there are various strategies to be explored for working with historical language varieties and small data (see, e.g., [Hedderich et al., 2021](#)), two of these are the use of **unsupervised** approaches and the use of techniques that require **less data**. These are the two requirements that motivated the choice of cohesion and conventionality, as unsupervised measures, in order to model literality, and these requirements will also act as constraints on the adaptation of the two chosen measures.

Historical corpora are in general much smaller than modern corpora, yet they are often more richly annotated. What are often considered ‘expensive’ resources for modern languages, such as manually constructed lexica and corpora with rich linguistic annotations, are more attainable than large amounts of text. Crucially, for historical varieties, corpora do not tend to grow: the data there is is what we have. While corpora for language stages after the widespread adoption of the printing press, such as the DTA corpus ([Berlin-Brandenburgischen Akademie der Wissenschaften, 2024](#)), spanning the 17th to early 20th centuries, can reach similar sizes to modern data sets – the complete DTA corpus contains 370 million tokens – this is not the case for older data sets. While often the bottleneck is the transcription and digitization of the manuscript sources, in other cases, there are simply few extant manuscripts to be digitized.

As our target historical language variety we con-

sider Middle High German, for which we consider two example corpora: the Reference Corpus of Middle High German (ReM, Roussel et al. (2024)) and the corpus of the Middle High German Conceptual Database (MHDBDB, Zepezauer-Wachauer (1992)). The former encompasses just over 2 million tokens and the latter just over 9 million. However, our vector representations must be trained on modern German corpora in order to evaluate against the annotated data set and in order to enable a comparison against the original measures using pre-trained models. We therefore simulate the low-data setting of the abovementioned historical corpora by the use of similarly sized modern corpora. As a stand-in for ReM, we use the “dev” and “train” portions of the Hamburg Dependency Treebank (Foth et al., 2014) at about 2 million tokens, and for MHDBDB we use the “2011 mixed” corpus of 1 million sentences from the Leipzig Corpora Collection (Goldhahn et al., 2012), which contains about 7.6 million tokens.

For these amounts of data, neural embeddings do not tend to provide the best results. Sahlgren and Lenci (2016) compared a range of different models of distributional semantics on different sizes of training corpora, and their study suggests that a count-based model transformed using PPMI and SVD could provide the best results with the amounts of data we have available. Such a model has a further advantage in that its operation is more transparent than a prediction-based one. Though, as the authors note, none of the models do particularly well in this setting, so it remains to be seen whether the measures will remain effective with these inputs.

In order to model a word’s use in a specific context (tokens), in addition to its distribution in the whole corpus (types), we adopt an approach to modelling specific usage contexts that is described in Geeraerts et al. (2023) and which ultimately goes back to Schütze (1998). Type vectors are constructed from word co-occurrences in the entire corpus (transformed with PPMI/SVD), then a token vector is constructed by adding together the type vectors for all of the context words that occur in a certain window around the target token to be represented. In effect, a token is represented as a set of second-order co-occurrences: Two tokens are similar when they co-occur with words that co-occur especially often.

We then implement cohesion and conventionality using either type vectors or token vectors as re-

quired. **Cohesion** (cf. Ehren, 2017) is defined here as follows: For each token instance w_i , we compare the type vector for w with the type vectors for all the words in the context of w_i , calculating the mean similarity between all pairs of these vectors both including and excluding the target expression w . If the mean similarity is greater without w , then this reflects lower lexical cohesion, and we expect it to correlate with less literality for this usage. Our adaptation of this measure differs from the original mainly in the embeddings used.

We take **conventionality** (cf. Socolof et al., 2022) to be defined as follows: A given word has a set of instances W , and conventionality is calculated for a single instance w_i by comparison with the other instances of this word $O = W \setminus w_i$. μ_O is the average token vector of the instances in O , and σ_O the component-wise standard deviation for these same instances. The conventionality is then:

$$\text{conv}(w_i) = \left\| \frac{w_i - \mu_O}{\sigma_O} \right\|_2 \quad (4)$$

This differs from the original formulation in that the original calculated the deviation of a particular phrase in which a word occurs versus all the other phrases in which the word also occurs, but we take a simplified approach. Since, in our low-data setting, the word in question is unlikely to have the same context more than once, we compare each instance against all of the other 49 instances for each lemma. Note also that the sign is reversed, since the scale in our evaluation data set uses higher numbers for less literal usages.¹

4 Experimental Results

For the 155 lemmas in the annotated data set, we calculated cohesion and conventionality with each of the three implementations for each of the ≈ 41.5 annotated instances, resulting in 38,556 instances total. Of these, 1357 instances were omitted, either because the lemmas do not occur in the background corpus or because there wasn’t sufficient context in the test sentence. For the remaining 37,199 combinations of sentence, lemma, measure, and implementation, we averaged the three raters’ judgments together in order to compare them with the given value.

For the sake of comparison, we also evaluate each of these measures in a setting as close as pos-

¹All of the code pertaining to these experiments is provided here under a free software license: <https://gitlab.rub.de/ajroussel/nlp4dh2024>.

sible to the original papers (*_orig). For cohesion, this means using pre-trained skip-gram word2vec vectors for the type vectors, and for conventionality, we use a pre-trained German BERT model to encode each sentence, from which we retrieve a contextualized representation for each target token instance.

The results of this comparison can be found in Table 1, and a visualization of the per-lemma correlations with average human judgments in Figure 1.

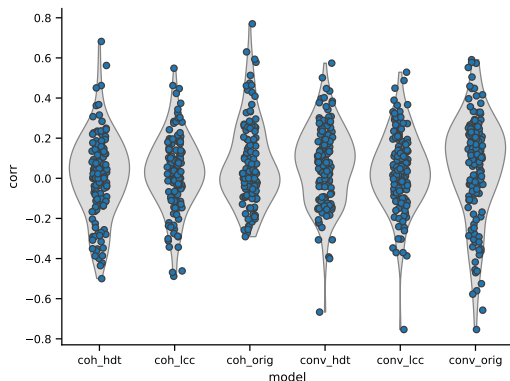


Figure 1: Correlations (Spearman’s ρ) between models and human judgments. Each dot represents the correlation for a particular lemma.

Table 1: Overall correlations for each measure and implementation. Impl. = Implementation, Desc. = Description, Coh. = Coherence, Conv. = Conventionality. “hdt” or “lcc” indicates the corpus used to create the vectors, and “orig” are the pre-trained vector representations.

Impl.	Desc.	Coh.	Conv.
hdt	tSVD, 2M	0.072***	0.073***
lcc	tSVD, 10M	0.050***	0.052***
orig	w2v/BERT	0.098***	0.010, ns

5 Discussion and Conclusion

Comparing the values of the two measures in their various implementations to the human judgments in the literal/non-literal data set does not appear to reveal any reliable patterns. Though in some cases the correlations are technically significant, the actual level of correlation is too low for either measure to be trusted in any particular case. The results suggest that these measures, implemented as described above, don’t correspond, in general, to the notion of literality that the annotators had in mind. As a result, it is also impossible to say whether

the adaptations of the measure for use with smaller corpora, such as for historical language varieties, were appropriate or whether they had any effect on the usefulness of the measures implemented.

An analysis of the correlations of the various measures with the judges’ ratings on a per-lemma basis was likewise inconclusive. As is evident in Figure 1, the correlations for particular lemmas can vary quite widely between strongly negative and positive correlations. We haven’t been able to find a clear reason for this; there are no apparent tendencies towards higher correlations when a lemma has a greater proportion of non-literal instances, for instance.

In future work, we plan to conduct more extensive annotation efforts specifically targeting literality in order to collect more fine-grained data to use in future experiments. Ideally, such a data set will cover not just particle verbs, but all open-class lexemes, and we plan to formulate detailed guidelines that will improve reproducibility and reusability of the data set.

6 Ethical considerations

Embeddings trained on corpus data scraped from the web, such as are employed in the comparison here, are known to have certain biases that could have had an effect on the outcomes of this study.

7 Limitations

We have characterized both of these measures as ones of ‘literality’ in general, but it’s still unclear to what degree (a) each of these individually or in combination correspond to a recognizable and coherent concept of literality, and (b) whether the conception of literality captured in the annotated data set corresponds to the aspects of literality that the measures relate to, or whether any apparent correlation is spurious. This study in its current form isn’t in a position to address these questions.

Acknowledgments

Thanks to the reviewers for their helpful feedback. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1475 – Projektnummer 441126958.

References

Berlin-Brandenburgischen Akademie der Wissenschaften. 2024. Deutsches Textarchiv: Grundlage

- für ein Referenzkorpus der neuhochdeutschen Sprache. <https://www.deutschestextarchiv.de/>. Accessed 2024-08-28.
- Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. **Unsupervised compositionality prediction of nominal compounds**. *Computational Linguistics*, 45(1):1–57.
- Rafael Ehren. 2017. **Literal or idiomatic? identifying the reading of single occurrences of German multiword expressions using word embeddings**. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–112, Valencia, Spain. Association for Computational Linguistics.
- Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. **Because size does matter: The Hamburg Dependency Treebank**. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang. 2023. *Lexical Variation and Change: A Distributional Semantic Approach*. Oxford University Press.
- Raymond W. Gibbs Jr., Darin L. Buchalter, Jessica F. Moise, and William T. Farrar IV. 1993. **Literal meaning and figurative language**. *Discourse Processes*, 16(4):387–403.
- Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. **Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages**. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).
- Michael A. Hedderich, Lukas Lange, Heike Adel, Janik Strötgen, and Dietrich Klakow. 2021. **A survey on recent approaches for natural language processing in low-resource scenarios**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.
- Maximilian Köper and Sabine Schulte im Walde. 2016. **Distinguishing literal and non-literal usage of German particle verbs**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics.
- George Lakoff. 1986. *The meanings of literal. Metaphor and Symbolic Activity*.
- Jeff Mitchell and Mirella Lapata. 2008. **Vector-based models of semantic composition**. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.
- Adam Roussel, Thomas Klein, Stefanie Dipper, Klaus-Peter Wegera, and Claudia Wich-Reif. 2024. *Referenzkorpus Mittelhochdeutsch (1050–1350)*. ISLRN 937-948-254-174-0.
- Magnus Sahlgren and Alessandro Lenci. 2016. **The effects of data size and frequency range on distributional semantic models**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980, Austin, Texas. Association for Computational Linguistics.
- Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. **A word embedding approach to predicting the compositionality of multiword expressions**. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.
- Agata Savary and Silvio Ricardo Cordeiro. 2017. **Literal readings of multiword expressions: as scarce as hen's teeth**. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 64–72, Prague, Czech Republic.
- Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. **Exploring vector space models to predict the compositionality of German noun–noun compounds**. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Hinrich Schütze. 1998. **Automatic word sense discrimination**. *Computational Linguistics*, 24(1):97–123.
- Michaela Socolof, Jackie Cheung, Michael Wagner, and Timothy O'Donnell. 2022. **Characterizing idioms: Conventionality and contingency**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4024–4037, Dublin, Ireland. Association for Computational Linguistics.
- Caroline Sporleder and Linlin Li. 2009. **Unsupervised recognition of literal and non-literal use of idiomatic expressions**. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.
- Julie Weeds, Thomas Kober, Jeremy Reffin, and David Weir. 2017. **When a red herring is not a red herring: Using compositional methods to detect non-compositional phrases**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short*

Papers, pages 529–534, Valencia, Spain. Association for Computational Linguistics.

Katharina Zeppezauer-Wachauer. 1992. Mittelhochdeutsche Begriffsdatenbank (MHDBDB). <https://mhdbdb.plus.ac.at/>. Universität Salzburg, accessed 2024-08-27.

Improving Latin Dependency Parsing by Combining Treebanks and Predictions

Hanna-Mari Kupari, Erik Henriksson, Veronika Laippala, Jenna Kanerva

TurkuNLP, University of Turku, Finland

{hanna-mari.kupari, erik.henriksson, mavela, jmnybl}@utu.fi

Abstract

This paper introduces new models designed to improve the morpho-syntactic parsing of the five largest Latin treebanks in the Universal Dependencies (UD) framework. First, using two state-of-the-art parsers, Trankit and Stanza, along with our custom UD tagger, we train new models on the five treebanks both individually and by combining them into novel merged datasets. We also test the models on the CIRCSE test set. In an additional experiment, we evaluate whether this set can be accurately tagged using the novel LASLA corpus (<https://github.com/CIRCSE/LASLA>). Second, we aim to improve the results by combining the predictions of different models through an atomic morphological feature voting system. The results of our two main experiments demonstrate significant improvements, particularly for the smaller treebanks, with LAS scores increasing by 16.10 and 11.85%-points for UDante and Perseus, respectively (Gamba and Zeman, 2023a). Additionally, the voting system for morphological features (FEATS) brings improvements, especially for the smaller Latin treebanks: Perseus 3.15% and CIRCSE 2.47%-points. Tagging the CIRCSE set with our custom model using the LASLA model improves POS 6.71 and FEATS 11.04%-points compared to our best-performing UD PROIEL model. Our results show that larger datasets and ensemble predictions can significantly improve performance.

1 Introduction

In recent years, significant progress has been made in morpho-syntactic dependency parsing for Latin, an advancement that greatly benefits a wide range of research in the humanities. Linguistically tagged corpora are crucial, as lemmatized corpora, for instance, are valuable also for historians searching for sources within databases. The Universal Dependencies (UD) framework plays a key role by organizing linguistic analysis into machine-readable

databases with columns in tab-separated value tables. These CoNLL-U formatted treebanks provide essential information on lemmas, parts of speech, morphological features, syntactic roles, and dependency relations. In the realm of Latin treebanks notable recent developments include the morphological harmonization of the five largest Latin treebanks (ITTB, LLCT, Perseus, PROIEL, and UDante¹), a significant milestone reached by Gamba and Zeman (2023a) as a continuation of earlier work on syntactic harmonization (Gamba and Zeman, 2023b).

Additionally, there have been many efforts to enhance the performance of Latin parsing tools. These include the EvaLatin campaigns Sprugnoli et al., 2022, 2024, as well as the application of GPT models for part-of-speech (POS) tagging (Stüssi and Ströbel, 2024). Despite these advancements, there remains potential for further improvement, particularly in syntactic parsing. For instance, the highest Labeled Attachment Score (LAS) reported by Gamba and Zeman (2023a) is 64.87% for the UDante and 59.43% for Perseus.

In the present study, we leverage the recently released harmonized treebanks (Gamba and Zeman, 2023a) to further enhance automatic parsing. Our focus is on the five largest established treebanks in the UD format, ensuring that our results are reliably comparable to previous studies. Our models can also easily be applied to parse new text corpora.

To achieve our goal, we employ two approaches: First, we train new parser models using these harmonized treebanks, along with two state-of-the-art parsers —Stanza (Qi et al., 2020) and Trankit (Nguyen et al., 2021)— as well as a custom UD tagger by fine-tuning a BERT-based Latin language model (Ströbel, 2022) following the architecture of Devlin et al. (2019). The parsing models are trained using both individual and diverse merged

¹<https://universaldependencies.org/la/>

treebanks.

Second, we investigate whether combining predictions from our newly trained models in a voting system targeting part-of-speech (POS) and morphological features (FEATS) tags improves performance. Our hypothesis is that selecting the most common prediction from the different models enhances the results in a 'majority vote wins' scenario.

Third, we use the voting setup of the different models to analyze how unanimous the various parser models are in their POS predictions. This provides insight into which tasks are accurately tagged and offers potential for identifying prevailing issues in the annotation guidelines.

Upon the publication of this paper,² all data, code, and results, as well as the models, will be made openly and freely accessible for non-commercial use. These resources include clear instructions, designed to be easily used by scholars who may not be familiar with language technology but wish to experiment with their own texts.

2 Previous work

The first Latin BERT model by [Bamman and Burns \(2020\)](#) provided the state-of-the-art POS scores of its time (Perseus 94.3%, PROIEL 98.2%, ITTB 98.8%). Similarly, [Nehrdich and Hellwig \(2022\)](#) reported very competitive LAS scores for the previous releases of the treebanks using a biaffine parser on top of a Latin BERT (ITTB 92.99%, PROIEL 86.34% and PERSEUS 80.16%).

There have been some trials with merging existing treebanks into larger training datasets. [Nehrdich and Hellwig \(2022\)](#) combined the ITTB, Perseus, and PROIEL treebanks, while [Smith et al. \(2018\)](#) trained a single model for all ancient languages, including three Latin treebanks. Additionally, [Kondratyuk and Straka \(2019\)](#) combined all the UD treebanks into a single multilingual dataset and trained a model for all UD languages. While these studies demonstrated the potential for improving performance by merging training data from multiple treebanks, the first reports only a single experiment, and the latter two do not focus specifically on Latin, leaving room for further experiments. The challenge of selecting and combining treebanks is also brought to attention in the latest EvaLatin Campaign ([Sprugnoli et al., 2024](#)).

Merging treebanks for training models has not

²<https://github.com/HannaKoo/Latin-Parsing>

been widely explored, likely because the developers of the treebanks have varied interpretations of the UD guidelines since the treebanks have been composed at different points in time (with continuous updates regarding the annotation guidelines). These discrepancies in annotations has complicated combining them into larger merged training datasets. The work of [Gamba and Zeman \(2023a\)](#) focuses on the harmonisation of the datasets, and they train models using only the individual treebanks.

Combining the predictions of several models through voting has been tested in many studies. E.g. early pioneering work by [Zeman and Žabokrtský \(2005\)](#) applied majority voting for four parsers for Czech, reporting improvements of 2%-points in dependency relation prediction. Combining parser outputs has also been used by [Passarotti and Dell'Orletta \(2010\)](#) to improve the parsing of the ITTB treebank. More recent work by [Stoeckel et al. \(2020\)](#) developed an ensemble classifier by applying a voting model on top of several POS taggers. Their voting model was designed to learn which predictions to trust in different contexts.

3 Data

There are five Latin UD treebanks used for training: the Index Thomisticus Treebank (ITTB) ([Passarotti, 2019](#)), the Late Latin Charter Treebank (LLCT) ([Cecchini et al., 2020b](#)), Perseus ([Bamman and Crane, 2011](#)), PROIEL ([Haug and Jøhndal, 2008](#)), and UDante ([Cecchini et al., 2020a](#)). For a concise numerical comparison of these Latin UD treebanks and a detailed description of their contents, see ³. For a general overview, see [Gamba and Zeman \(2023b\)](#).

The efforts of [Gamba and Zeman \(2023a\)](#) are crucial for merging the treebanks and serve as a foundation of our model training. These harmonized treebanks are accessible at a GitHub repository ⁴. For a concise numerical overview and a brief description of the treebanks used in this study, refer to Table 1.

³<https://universaldependencies.org/treebanks/la-comparison.html>

⁴https://github.com/fjambe/Latin-variability/tree/main/morpho_harmonization/morpho-harmonized-treebanks

3.1 CIRCSE test set

The novel sixth UD Latin treebank, CIRCSE⁵, consists solely of a test set because of its small size along the UD guidelines. This test set is valuable for evaluating our models because it differs from the established larger treebanks, which predominantly feature texts from the middle ages. For instance, the ITTB and LLCT together contain 692K tokens, whereas Perseus focuses on Classical texts with a total of only 29K tokens. CIRCSE is also distinct in genre, featuring a total of 13,294 tokens of tragedy: *Hercules Furens* (7,714 tokens, 555 sentences) and *Agamemnon* (5,580 tokens, 409 sentences) by Seneca (c. 4 BC – AD 65), along with the treatise *Germania* (5,674 tokens, 299 sentences) by Tacitus (c. AD 56 – c. 120).

3.2 Merged treebanks

Merging treebanks presents challenges not only due to potential differences in annotation guidelines but also because of the linguistic variation they reflect. The five treebanks span several millennia and cover a wide range of genres, factors that can influence the performance of models trained on them. One of the key research questions we explore is whether, for example, the inclusion of a large amount of medieval Latin training data affects the parsing results for Classical Latin.

In addition to merging all the training datasets, we combine the individual treebanks into five thematically organized merged treebanks, as shown in Table ??, based on a holistic understanding of the nature of the different Latin UD datasets. We also experiment with merged sets focused on specific time periods, drawing on a heuristic understanding of historical linguistics and the evolution of the Latin language. The goal is to compile sets that support one another, rather than confuse the models with training data that is too varied or even contradictory. Beyond linguistic considerations, to address machine learning challenges and mitigate the risk of overfitting—particularly when working with datasets from unequally sized and heterogeneous treebanks—the merged training sets were constructed by iteratively concatenating one-fifth of each individual treebank, ordered from smallest to largest, into the new datasets.

⁵https://universaldependencies.org/treebanks/la_circse/index.html

3.3 The Corpus Corporum monolingual training set

While most of our experiments are based on the widely applied Stanza and Trankit parsers (see Section 4), neither of them support using a dedicated pre-trained Latin language model. Therefore, we also experiment with our custom tagger utilizing a language model trained on Latin data only (see Section 4.3). The language model (Ströbel, 2022) has been produced by using the Corpus Corporum dataset (Roelli, 2014). This dataset contains a considerably large portion of patristic texts from the Patrologia Latina (8.4 M words). For a concise overview of the texts currently included in this database see the listing on the project website⁶. The previous work of Bamman and Burns (2020) with a monolingual model for POS tagging is produced with a very large dataset of 642.7M tokens that includes for example Latin Wikipedia of 16M tokens. This provides obvious problems as to reliable quality of the training data, since contributions to Vicipaedia are not subject to expert language check and the RoBERTa Latin model by Ströbel (2022) is focused to solve this very issue.

3.4 The LASLA dataset

Since texts from the Classical period are underrepresented in the UD treebanks, we conduct a small experiment using the non-UD LASLA dataset, which lacks dependency parsing annotation. In terms of POS tagging, lemmatization and morphology, the 1.8M-token LASLA dataset is notably large, created through a joint effort by members of the LiLa and LASLA teams.⁷ We use the LASLA corpus as a basis to make our own train, dev, and test sets for a small-scale experiment aimed at improving our custom model for the POS and morphological analysis of the CIRCSE test set. Our modification of the CoNLL-U Plus formatted files excludes the texts in the CIRCSE test set (see 3.1) and removes non-relevant fields. The larger files are split and concatenated in random order.

4 Methods

In our aim to improve morpho-syntactic parsing tools for Latin, we use two different methods: training new models and experimenting with a voting system. Our first task is the training of new parser

⁶<https://mlat.uzh.ch/browser?path=/>

⁷<https://github.com/CIRCSE/LASLA?tab=readme-ov-file>

Token counts or words in datasets					
Dataset	Short Description	Train	Dev	Test	Total
CIRCSE	Seneca’s tragedies and Tacitus’ treatise	-	-	19 483	19 483
ITTB	Texts of Thomas of Aquinas, 13th century	392 017	29 968	29 920	451 905
LLCT	8th century legal charters from Tuscany	194 193	24 195	24 079	242 467
Perseus	Classical auctors e.g. Caesar and Ovid	16 859	1 566	11 149	29 574
PROIEL	Classical auctors and New Testament	172 261	13 955	14 114	200 330
• Classical	E.g. Cicero and Palladius	76 647	-	-	76 647
• Vulgate	Jerome’s Vulgate	95 614	7 123	-	102 737
UDante	Works of Dante Alighieri, 13th-14th century	30 567	11 689	13 502	55 758
CC	Massive Corpus Corporum text database				162 M
LASLA	Classical Latin database	1 856 296	32 756	35413	1 856 296

Table 1: Overview of the used datasets for train, dev and test. We have split PROIEL to include Classical secular texts and Vulgate. For Perseus, where the original release does not include a separate development set for parameter optimization, we created one by dividing the train set. The UD CIRCSE treebank only contains a test set due to its size. The Corpus Corporum dataset is the basis for the monolingual BERT (Ströbel, 2022) used for our custom model UD tagger. Our modification of the LASLA database (<https://github.com/CIRCSE/LASLA/tree/main>) is used in an experiment to improve the results of the CIRCSE test set.

Training data	ITTB	LLCT	Perseus	PROIEL	UDante	Tokens in total
Classical Latin			9%	91%		205 K
Late and Medieval Latin	62%	32%			6%	683 K
Later and Christian Latin	54%	28%		13%	5%	785 K
Merged	48%	25%	2%	21%	5%	887 K

Table 2: Overview of the merged treebanks used for training Stanza and Trankit and fine-tuning the custom model.

models based on the newest treebanks described in Tables 1 and 2. For full morpho-syntactic parsing, we apply the commonly used Trankit (Nguyen et al., 2021) and Stanza (Qi et al., 2020) toolkits. As neither Trankit nor Stanza support the usage of a custom pretrained language model, we also experiment with a custom part-of-speech and morphological tagger trained on top of a monolingual Latin language model (Ströbel, 2022) following the task-specific fine-tuning of Devlin et al. (2019).

4.1 Trankit

Trankit (Nguyen et al., 2021) is a light-weight transformer based toolkit, which provides a trainable pipeline for morpho-syntactic parsing. It reports outperforming prior multilingual NLP pipelines over sentence segmentation, POS and FEAT tagging as well as in dependency parsing while maintaining competitive performance for tokenization, multi-word token expansion, and lemmatization over 90 UD treebanks. It is based on training adapter modules (Houlsby et al., 2019; Pfeiffer et al., 2020) on top of the multilingual pretrained XLM-R language model (Conneau et al., 2020).

The parser is designed to be efficient in multilingual usage (shared multilingual language model), while still giving state-of-the-art results for individual treebanks (treebank-specific adaptors).

4.2 Stanza

Stanza (Qi et al., 2020) is a trainable, language-agnostic neural pipeline for morpho-syntactic parsing. Stanza includes a Bi-LSTM encoder capable of utilizing pre-trained word embeddings, and uses the biaffine neural dependency parser by Dozat and Manning (2017). This is the same parser that Gamba and Zeman (2023a) employed. We use standard model training in order to have a model that matches the Trankit training to ensure a reliable comparison between the models.

4.3 Custom tagger with a Latin language model

Earlier studies, e.g. Pyysalo et al. (2021); Bamman and Burns (2020), have shown that for certain languages the usage of a dedicated monolingual language model may result in better performance compared to multilingual models or not using a

pretrained language model at all. While neither Trankit nor Stanza support the usage of a custom pretrained language model, we implement a POS and morphological tagger by fine-tuning a monolingual Latin language model. As a pretrained language model we use the `pstroe/roberta-base-latin-v3`⁸ pretrained on the Corpus Corporum Latin text collection (see Section 3.3). The tagger jointly predicts the POS and morphological features by adding a task-specific token classification layer on top of the pretrained language model, following the architecture of [Devlin et al. \(2019\)](#). The classification layer is trained on treebank data updating also the weights of the original language model.

4.4 Voting

In POS tag and FEATS predictions voting we run a simple majority vote of the three parsers (Trankit, Stanza, and Custom tagger), for each treebank selecting the generally best performing model of each parser. In a tie situation, the voting defaults to Trankit which generally receives the best individual scores. The voting script does not take into account the fact that the numerically highest scores for POS and UFEATS might come from different models, and our preference is for overall best results.

For POS tags, the possible voting scenarios when using three parsers are cases where all three agree, two outvote the third one and all parsers disagree. When analysing the model predictions for the Perseus treebank, in 86% of tokens the three parsers agree on UPOS, in 13% of tokens there is a majority agreement, and only in a bit more than 1% all three parsers disagree on UPOS.

However, in terms of morphological features the same agreement rates on Perseus are 59%, 31%, and 10% respectively, when voting on the level of full feature analyses — the entire FEATS field that consists of several categories such as number and tense. The large variation in predicted feature combinations therefore increases the percentage of tokens where there is no majority consensus available (10%).

To be able to at least partially account for these tokens as well, for morphological features we proceed the voting in two steps. First, the voting is done on the level of full feature analysis (e.g. for nouns this means that all the diverse elements in the category, such as case, number and gender), but

⁸<https://huggingface.co/pstroe/roberta-base-latin-cased3>

in cases where we are not able to find a majority vote, we continue to the second option of voting on category level. In the second step, the feature analyses are split into individual (category, value) -pairs, and for each category we run the majority voting of values predicted for that particular category. To avoid the situation where the final analysis is a union of different categories predicted by three parsers, we obtain the categories from the default Trankit parser, therefore in practice only voting values for Trankit predicted categories. It should also be noted that the LASLA model for CIRCSE is not included in the vote, as it would require a close reading of potentially non-UD-style morphological annotations, which the script does not consider.

5 Results

The performance of the trained models is summarized in Table 3, which presents the results for the five largest established treebanks. Additionally, the outcomes specific to the CIRCSE treebank are detailed in Table 6 and Table 7. The findings underscore the importance of selecting optimal treebanks for training, as discussed by [Sprugnoli et al. \(2024\)](#). While the prevailing trend in training large language models has been to utilize increasingly larger datasets, our results indicate a different effect. Specifically, the Perseus treebank shows significant improvement when trained with the Classical dataset, indicating that quality of data is more critical than quantity, challenging the assumption that "more is better". The effects of this improvement are highlighted in Table 8.

The complete set of metrics is available on the project's GitHub page⁹ and the all CoNLL-U formatted treebanks respectively¹⁰. In this paper, we report and discuss the scores for tokenization, POS, morphological features (FEATS), lemmatization, and syntax, including both the unlabeled attachment score (UAS) and labeled attachment score (LAS). For the custom tagger, only the UPOS and FEATS results are relevant. All metrics were generated using the UD evaluation tools, based on the CoNLL 2018 shared task script¹¹.

In the results presented below we discuss the

⁹https://anonymous.4open.science/r/Latin-Parsing-627B/Results/Evaluation_metrics/eval_table.tsv

¹⁰https://anonymous.4open.science/r/Latin-Parsing-43B5/Results/conllu_files

¹¹<https://github.com/UniversalDependencies/tools/blob/master/eval.py>

Compilation of Results	Tasks:					
	Tokens	UPOS	UFeats	Lemmas	UAS	LAS
ITTB						
Stanza	100.00	98.64	96.16	99.05	88.50	86.61
Trankit	99.99	98.99	97.52	97.63	92.09	90.71
Trankit Late and Christian	100.00	99.05	97.61	97.87	91.86	90.52
Trankit Five Merged	99.99	99.07	97.55	97.82	91.90	90.41
Custom tagger Late and Christian	-	98.72	96.61	-	-	-
LLCT						
Stanza	100.00	99.61	96.95	98.07	95.85	94.83
Trankit	99.99	99.66	97.36	96.50	96.15	95.37
Trankit Late and Medieval	99.99	99.66	97.18	96.69	96.46	95.51
Custom tagger	-	99.14	95.67	-	-	-
Perseus						
Stanza	99.94	89.44	80.17	80.97	69.75	61.93
Stanza Classical	99.92	90.09	81.33	85.89	75.28	68.29
Trankit Classical	99.74	90.50	83.25	74.60	77.89	71.28
Trankit Five Merged	99.79	91.83	80.94	76.55	77.72	70.59
Custom tagger Classical	-	89.58	82.58	-	-	-
Custom tagger Five Merged	-	89.66	78.43	-	-	-
PROIEL						
Stanza	99.99	97.22	92.14	96.63	78.12	74.56
Trankit	99.87	97.29	92.77	89.37	84.09	80.97
Trankit Five Merged	99.88	97.30	92.96	89.24	83.94	80.92
Custom tagger Five Merged	-	96.44	91.64	-	-	-
UDante						
Stanza	99.65	89.98	81.00	86.94	68.37	59.15
Trankit Five Merged	99.66	91.46	84.42	77.50	79.63	73.42
Custom tagger Five Merged	-	89.91	82.24	-	-	-

Table 3: A compilation of the most important F1-scores. The best score for each treebanks is in bold.

most relevant numbers and some case study examples. In Table 9 we also include the previous state-of-the-art outcomes from two recent studies. Our state-of-the-art results demonstrate improvements in POS-tagging of 8.41 %-points for Perseus, 7.78 for PROIEL, and 5.93 for UDante compared to the findings of Stüssi and Ströbel (2024). Additionally, our results show an improvement in LAS of 11.85%-points for Perseus and 16.10%-points for UDante compared to Gamba and Zeman (2023a).

All numerically highest F1 scores achieved by the models are in the Table 3. The effects of the merging of training data set for training are in Table 8. The results of the majority vote win for POS and FEATS are in Table 5.

5.1 Tokenization

Tokenization results have very little room for improvement, the best models already obtaining an F1

score of 100 % for ITTB and LLCT with individual training. From close reading we find that the only aspect of tokenization that requires improvement is the prediction of multi-word tokens (MWTs). This issue arises from the complete absence or inclusion of only a few trivial MWTs in these corpora. E.g. the ITTB train set contains only instances of *nonne* 'isn't it?', which is clearly insufficient for effectively training the models on something as complex as Latin enclitics). Upon close reading the output, we identified predictions that are significantly off. For instance, in the Perseus corpus parsed by Stanza, the word *pulsabantque* 'and beat' is incorrectly tokenized as "*pullaaa*" and "*que*" instead of the correct "*pulsabant*" and "*que*,"

The tokenization of the CIRCSE test set achieved a perfect accuracy of 100.00% with the Stanza PROIEL model. However, this test set lacks punctuation, leading to poor performance in the

task of **sentence segmentation** across all models. Several of our models were unable to segment sentences and attempted to dependency parse the entire dataset as a single 19K words long sentence. To address this, we experimented using a crude fix of adding a full stop at the end of each sentence using a script, and assigned a mock HEAD-tag pointing to the last word of each sentence, resembling the use of GS segmentation. For further details and results of this experiment, see Table 7.

5.2 Part-of-Speech (POS)

Overall, the results for POS tagging have for a long time highly accurate and for most treebanks can only be marginally improved.

All the results of the POS vote are written in a new ConLL-U-styled tsv-table that first includes the winner of the majority vote, the predicted forms in the following order: Trankit, Stanza and custom model.¹² After that a column indicates the results of the vote being either unanimous, two-to-one or even. The resulting file¹³ includes also a column that indicates if the result of the vote is correct, this information is especially informative for close reading. Scholars are able to form a general idea of what kind of tasks the parsers are capable of predicting and can especially focus on the difficulties and understand if there is an underlying trend that could be fixed (i.e. relating to the annotation guidelines).

The most interesting cases are the ones with dispersed results and here we will highlight some case examples. From the ITTB treebank we find a case with the word *nesesse-esse* 'necessarily existent' with POS predictions: ADJ, VERB, AUX where our custom model gets it right according to the GS of the morphological harmonization, but the earlier realise tags this as NOUN. From LLCT we find instances like *decimas* (from phrase *per quadraginta annos abuerunt consuetudo offertas et decimas dare ad predicta ecclesia*) as ADJ, NOUN, NUM, where Stanza gets the POS tag of 'tithes' right. There are a lot of expressions of date, for example *in mense december* where one instance the vote is even for *december* resulting as NOUN, ADJ, NUM while all other instances in the test dataset get it unanimously right as ADJ. From the expression *adfinis terra* 'boundaries of the land' *adfinis* as ADJ, NOUN, ADP when Trankit gets it correct. The Perseus and UDante outputs have substantially

¹²Results/conllu_files/voted_extended

¹³Results/conllu_files/gold_extended

PROIEL	Correct	%	Wrong	%
Unanimous	13 295	99%	132	0.98%
Two to one	463	75%	154	25%
Dispersed	15	44%	19	55%

Table 4: An example of the accuracies of the voting on POS tagging in PROIEL

more even votes than the other five established treebanks. These include *iuro* NOUN, VERB, ADJ from the phrase *per flumina iuro* (swear by the rivers), also we find *Aeoliis* as ADJ, PROP, ADP where Trankit gets it correct. From PROIEL *promissa* as NOUN, ADJ, VERB from the expression *ceterorum que promissa* which is easy to understand, since the participle *promissum* 'promises' and we would also imagine this being difficult for Latin students, but Trankit is correct with NOUN. From UDante the phrase *praedictis finibus* 'of the aforementioned borders' where the participle is predicted as DET, VERB, ADJ and only Stanza is correct. An example of voting accuracy in PROIEL in Table 4.

For **CIRCSE**, the best UD framework based model part-of-speech tagging result comes from Stanza trained on PROIEL at 84.46%, but other models are close. However, our small experiment with the LASLA model does bring an improvement of 6.71%-points (UPOS **91.17%**) hinting that the results for many other out of genre texts from the Classical period might be considerably improved with larger training data.

5.3 UFeats

The morphological analysis results seem to vary greatly between different treebanks, from **ITTB** reaches already a very impressive result of **97.61%** but for **UDante** only at **84.42%**. This seems to follow the trend, that when there is enough of in-domain training data, the results have very little room for improvement. The best UD framework based **CIRCSE** morphological analysis is achieved with the Stanza PROIEL model (59.48%) as was for POS. Surprisingly using the LASLA model gives an improvement of 11.04%-points (UFeats **70.52%**).

5.4 Lemmas

Accurate automatic lemmatization is a very relevant task for a highly inflected language like Latin. The results have a high amount of variation across different treebanks but overall Stanza models seem to consistently outperform on this task. The re-

sults for **ITTB** comes from the Stanza individually trained model at an impressive **99.05%** as well as for **LLCT 98.07%**, **PROIEL 89.24%** and **UDante 86.94%**. For **Perseus** the best score **85.89%** is produced by using the Stanza Classical model.

The best lemmatization score for **CIRCSE** is Stanza Five Merged **78.00%**.

5.5 Unlabeled Attachment Score: UAS

Latin papers on automatic parsing usually report the unlabelled attachment scores (UAS) along with labeled attachment score (LAS). The UAS metric means the percentage of words that are assigned the correct head in the sentence. The results syntactic tagging vary greatly. The **Perseus** treebank benefits from only seeing training data from its own time period. On the contrary, the same does not apply for **UDante**, which benefits greatly from the merged training data and obtains a **79.63%** score with Trankit Five Merged (66.79% on UDante).

For **CIRCSE** the best score is only **51.29%** by the Trankit Five Merged model, this is understandable considering how far the training model data is as genre for parsing the tragedies. Adding the punctuation with a very coarse simple full stop addition at the end of each sentence makes this dataset much easier for models to syntactically parse, this alone leads to a 59.16% with above mentioned model.

5.6 Labeled Attachment Score: LAS

For the second metric on syntax, the Labeled Attachment Score LAS, the results are in line with UAS findings. The LAS score is the percentage of words that are assigned both to the correct head and the correct dependency label. The results in Table 3 show that the results tend to be dependent on the amount of similar training data.

The LAS score of the **CIRCSE** test set shows the true nature and difficulty of out of domain Latin syntax parsing. Our experiment reflects the more of a real life situation with parsing new data and our best score is **44.54%** from Trankit Five Merged. The altered punctuation yields a 50.91% score on same model. The EvaLatin2024 (Sprugnoli et al., 2024) results reach 77.41% for prose and 75.75% poetry. The task performance is not comparable since for the shared task included the use of train and dev datasets and had only the dependency parsing task. Straka et al. (2024) report leveraging the GS morphological annotation as an additional input for the parser.

5.7 Voting results

The results of the voting experiment are reported in Table 5, giving the baseline scores for the three parsers (Trankit, Stanza, custom tagger), and the majority voting results. In addition to this, we also report *Oracle* score to illustrate the theoretical upper bound for voting when it is based on these three parsing models, i.e. the accuracy of a hypothetical voting system that is always able to select the best option among the predictions. Based on the results by Zeman and Žabokrtský (2005) we expected a possible an increase of roughly two percentage points. The improvement of the voting results is reported in 5 and ranges from 0.00% to +0.89 for POS tagging and for FEATS from +0.09% to +3.15%.

6 Conclusion and future studies

The task of full morpho-syntactic parsing across the five largest established treebanks comprises 30 subtasks, of which 8 are best performed by the Trankit Five Merged model. This model demonstrates particular strength in part-of-speech labeling. Additionally, Stanza’s lemmatization capabilities are noteworthy, consistently achieving the highest numerical values across all five treebanks.

Overall it can be stated that merging the available five Latin UD datasets is very beneficial especially when it comes to smaller treebanks and out of domain parsing. With our experiments, by using thematically compiled and everything merged datasets, we are able to set a new state of the art for many morpho-syntactic parsing tasks. The average improvement of our final results are reported in Table 9. Our initial results of morphological features are even further improved by using the FEATS atomic voting system especially on the smaller treebanks. The results reaching +3.15 %-points.

Future studies should first focus on addressing the issues related to the treatment of multi-word tokens. One approach could involve ensuring that the five established treebanks strictly adhere to current guidelines, such as avoiding the splitting of enclitics (e.g., *-que* 'and') into separate tokens. Additionally, the introduced voting system could be further refined and applied to a gold-standard pre-tokenized input, followed by a detailed numerical error analysis and close reading. This enables determining the specific morphological annotation tasks that our current models succeed upon. Such analysis could also determine whether observed errors

	ITTb		LLCT		Perseus		PROIEL		UDante		CIRCSE	
	UPOS	UFeat	UPOS	UFeat	UPOS	UFeat	UPOS	UFeat	UPOS	UFeat	UPOS	UFeat
Trankit	99.07	97.55	99.63	97.15	91.83	80.94	97.30	92.96	91.46	84.42	83.21	57.76
Stanza	98.64	96.15	99.61	96.96	90.81	82.03	97.14	92.18	89.85	80.92	84.47	56.85
Custom	98.72	96.61	99.14	95.67	89.58	82.58	96.44	91.64	89.91	82.24	79.72	55.29
Majority	99.07	97.64	99.64	97.32	92.72	85.73	97.78	93.98	91.73	85.25	85.25	60.23
Change	+0.00	+0.09	+0.01	+0.17	+0.89	+3.15	+0.48	+1.02	+0.27	+0.83	+0.78	+2.47
Oracle	99.60	99.01	99.82	98.46	96.11	92.64	98.83	96.98	94.19	90.69	90.22	65.31

Table 5: Results of the majority voting system compared to the three individual models used in voting. *Oracle* stands for a theoretical upper bound for voting of always selecting the best option among the predictions.

suggest the need for further harmonization of the treebanks themselves or are these cases difficult to grammatically analyze as such?

On one hand, many tasks are successfully accomplished using a single treebank for training, development, and testing, as demonstrated by the ITTB data, which does not require the inclusion of additional treebanks for improving performance. This highlights the importance of incorporating new genres across a broad time span into the UD Latin treebank family, ensuring that the training data is sufficiently diverse, comprehensive and large enough. While the development of novel gold-standard annotated datasets offers significant benefits, it is also highly demanding in terms of human resources. We hope that our high-performing models will facilitate the annotation of these datasets by providing accurate predictions that serve as a strong starting point for manual corrections, thereby easing the process.

On the other hand, one of the conclusions drawn from our diverse merged training sets is that the notion of "Latin is Latin" does not hold true. It is well established that medieval Latin is distinctly different from Classical Latin. In practical terms most scholars often identify themselves as experts in one or the other. However, a possible future study could investigate the specific attributes in a treebank's training data that make a parser model particularly adept at Classical or medieval Latin.

Another conclusion from our experiments is that the accuracy of parsing Latin from the Classical period (broadly defined) is diminished when the model is exposed to medieval training data. This warrants further exploration to define the characteristics that distinguish the two and will shed more light into computational historical linguistics. One study could be the evolution of medieval Latin and the extent to which medieval treebanks reflect preserving features of Classical Latin, analyzed by auctor and decade. It might reveal how well and what

ways medieval writers were competent in Classical Latin. Another potential research direction is to investigate why parsing the UDante treebank appears less selective, with all five merged models performing well. This raises the question of whether users of Latin from this late medieval period were equally accustomed and influenced by reading both Classical and medieval authors. Alternatively, this phenomenon might be explained by the size of the training data, where additional examples contribute to improved results, as our LASLA experiment in the CIRCSE test set show.

7 Limitations

Firstly, the harmonization of UD Latin syntactic annotation (Gamba and Zeman, 2023b) and morphological annotation (Gamba and Zeman, 2023a) has been taken as a given and we have not subjected the annotations to any closer examination. As suggested by the case study sample finding of *nesesse-esse* 'necessarily existent' (as discussed in the Section 5.2) the training datasets might include seldom errors from automatic processing. Secondly, the data in the LASLA corpus¹⁴ has not been examined for any potential divergences from the UD framework. We don't inspect the results from the reserved test set we have set aside for possible further experiments on the LASLA corpus based model with our custom model. This would need more resources and we leave this for the future, since our focus only on one experiment of the CIRCSE test set.

Acknowledgements

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources as well as The Emil Aaltonen Foundation for grant "Exploring linguistic variation in medieval Latin using computational methods" for Hanna-Mari Kupari 2022-2024.

¹⁴<https://github.com/CIRCSE/LASLA/tree/main>

References

- Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- David Bamman and Patrick J. Burns. 2020. [Latin bert: A contextual language model for classical philology](#). *Preprint*, arXiv:2009.10053.
- David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Flavio M Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020a. Udante: First steps towards the universal dependencies treebank of dante’s latin works. Accademia University Press.
- Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020b. [A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942, Marseille, France. European Language Resources Association.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Timothy Dozat and Christopher D. Manning. 2017. [Deep biaffine attention for neural dependency parsing](#). In *International Conference on Learning Representations (ICLR)*.
- Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA.
- Federica Gamba and Daniel Zeman. 2023b. Universalising latin universal dependencies: a harmonisation of latin treebanks in ud. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dag Trygve Truslew Haug and Marius L. Jøhndal. 2008. [Creating a parallel treebank of the old indo-european bibletranslations](#).
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.
- Dan Kondratyuk and Milan Straka. 2019. [75 languages, 1 model: Parsing Universal Dependencies universally](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.
- Sebastian Nehrlich and Oliver Hellwig. 2022. [Accurate dependency parsing and tagging of Latin](#). In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 20–25, Marseille, France. European Language Resources Association.
- Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. [Trankit: A lightweight transformer-based toolkit for multilingual natural language processing](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.
- Marco Passarotti. 2019. [The Project of the Index Thomisticus Treebank](#), pages 299–320. De Gruyter Saur, Berlin, Boston.
- Marco Passarotti and Felice Dell’Orletta. 2010. [Improvements in parsing the index Thomisticus treebank. revision, combination and a feature model for medieval Latin](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.
- Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. [WikiBERT models: Deep transfer learning for many languages](#). In *Proceedings*

- of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 1–10, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv.org*.
- Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.
- Philipp Roelli. 2014. The corpus corporum, a new open latin text repository and tool. *Archivum Latinitatis Medii Aevi*, 72(1):289–304.
- Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.
- Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.
- Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.
- Manuel Stoeckel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. 2020. Voting for POS tagging of Latin texts: Using the flair of FLAIR to better ensemble classifiers by example of Latin. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135, Marseille, France. European Language Resources Association (ELRA).
- Milan Straka, Jana Straková, and Federica Gamba. 2024. úfal latinpipe at evalatin 2024: Morphosyntactic analysis of latin. *Preprint*, arXiv:2404.05839.
- Phillip Benjamin Ströbel. 2022. Roberta base latin cased v2.
- Elina Stüssi and Phillip Ströbel. 2024. Part-of-speech tagging of 16th-century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206, St. Julians, Malta. Association for Computational Linguistics.
- Daniel Zeman and Zdeněk Žabokrtský. 2005. Improving parsing accuracy by combining diverse dependency parsers. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 171–178, Vancouver, British Columbia. Association for Computational Linguistics.

A Appendix

CIRCSE test set results	Tasks:					
Model Name	Tokens	UPOS	UFeats	Lemmas	UAS	LAS
Stanza PROIEL	100.00	84.46	59.48	72.37	48.18	41.38
Trankit PROIEL	99.24	81.50	55.39	60.08	49.44	41.92
Custom Perseus	-	76.29	47.79	-	-	-
Custom PROIEL	-	79.72	55.29	-	-	-
Custom Five Merged	-	81.30	57.11	-	-	-
Custom Classical	-	80.84	56.53	-	-	-
LASLA	-	91.17	70.52	-	-	-
Stanza Classical	100.00	84.37	56.79	73.36	49.64	43.03
Stanza Five Merged	99.98	82.56	51.23	78.00	47.00	40.14
Trankit Classical	99.71	83.08	57.09	62.87	50.57	43.06
Trankit Five Merged	99.82	83.21	57.76	68.15	51.29	44.54

Table 6: The results of the CIRCSE test set. For models trained on individual treebank data only the results for PROIEL are given for all models, since both Stanza and Trankit Perseus models failed to run because of severe sentence segmentation issues.

CIRCSE altered test set	Tasks					
Automatically added punctuation	Tokens	UPOS	UFeats	Lemmas	UAS	LAS
<i>Stanza ITTB</i>	<i>99.98</i>	<i>81.64</i>	<i>56.32</i>	<i>73.32</i>	<i>50.49</i>	<i>41.53</i>
<i>Stanza LLCT</i>	<i>99.99</i>	<i>75.41</i>	<i>40.54</i>	<i>56.13</i>	<i>37.24</i>	<i>25.18</i>
<i>Stanza PROIEL</i>	<i>100.00</i>	<i>79.98</i>	<i>62.06</i>	<i>74.20</i>	<i>46.17</i>	<i>38.59</i>
<i>Stanza Perseus</i>	<i>99.93</i>	<i>83.96</i>	<i>57.26</i>	<i>70.16</i>	<i>46.75</i>	<i>38.43</i>
<i>Stanza Classical</i>	<i>100.00</i>	<i>85.81</i>	<i>59.54</i>	<i>75.46</i>	<i>54.20</i>	<i>46.93</i>
<i>Stanza Five Merged</i>	<i>100.00</i>	<i>83.94</i>	<i>54.33</i>	<i>79.58</i>	<i>53.89</i>	<i>46.60</i>
<i>Trankit Classical</i>	<i>99.78</i>	<i>85.21</i>	<i>59.75</i>	<i>65.44</i>	<i>56.61</i>	<i>48.43</i>
<i>Trankit Late and Christian</i>	<i>99.80</i>	<i>84.85</i>	<i>58.20</i>	<i>66.69</i>	<i>54.53</i>	<i>45.41</i>
<i>Trankit Late and Medieval</i>	<i>99.74</i>	<i>82.68</i>	<i>55.35</i>	<i>63.18</i>	<i>51.99</i>	<i>42.52</i>
<i>Trankit Five Merged</i>	<i>99.79</i>	<i>87.05</i>	<i>61.39</i>	<i>71.73</i>	<i>59.16</i>	<i>50.91</i>

Table 7: The effects to the performance of the different models with the added punctuation to the CIRCSE gold standard test set. The results are not comparable to the UD released test set and given in italics.

Effects of merged treebanks in training	Tasks:					
Treebank and model	Tokens	UPOS	UFeats	Lemmas	UAS	LAS
ITTB						
Custom tagger	-	98.66	96.50	-	-	-
Improvement from Late and Christian	-	0.06	0.11	-	-	-
LLCT						
Trankit	99.99	99.66	97.36	96.50	96.15	95.37
Improvement from Late and Medieval	0.00	0.00	-0.18	0.19	0.31	0.14
Perseus						
Stanza	99.94	89.44	80.17	80.97	69.75	61.93
Improvement from Classical	0.02	0.65	0.96	4.92	5.53	6.36
Trankit	99.46	88.90	77.98	63.99	74.08	66.97
Improvement from Classical	0.28	1.60	5.27	10.61	3.81	4.31
Custom tagger	-	86.29	76.17	-	-	-
Improvement from Classical	-	3.29	6.41	-	-	-
PROIEL						
Custom tagger	-	96.42	91.26	-	-	-
Improvement from Five Merged	-	0.02	0.38	-	-	-
UDante						
Trankit	99.50	91.17	80.71	73.89	75.92	68.65
Improvement from Five Merged	0.16	0.29	3.71	3.61	3.71	4.77
Custom tagger	-	87.43	75.84	-	-	-
Improvement from Five Merged	-	2.48	6.40	-	-	-
Average improvement	0.15	1.19	2.78	4.84	3.34	3.90

Table 8: The most important results of the merging of diverse training data.

Tasks:	POS		UFEATS		UAS		LAS	
Treebank	Our highest	Change	Our highest	Change	Our highest	Change	Our highest	Change
ITTB	99.07	4.19	97.64	1.49	92.09	-0.19	90.71	2.42
LLCT	99.66	5.16	97.36	0.55	96.46	0.38	95.51	0.60
PERSEUS	92.72	8.41	85.73	7.87	77.89	8.92	71.28	11.85
PROIEL	97.78	7.78	93.98	1.26	84.09	-0.82	80.97	-0.28
UDante	91.73	5.93	85.25	5.95	79.63	12.84	80.97	16.10
Average change		6.29		3.42		4.23		6.14

Table 9: Summary of our best F1 scores. The ones produced by the voting system are given in a bold typeset. The change as percentage points to the most recent POS tagging study by [Stüssi and Ströbel \(2024\)](#). For **ITTB** the best score **99.07%** is predicted by Trankit Five Merged (in experimenting with a GPT model on POS tagging the best results reported by [Stüssi and Ströbel \(2024\)](#) is 94.88 produced on GPT-4 train1000). The same applies for **Perseus** as well **91.83%** (84.31 on GPT-4 train2000), **PROIEL** at **97.30%** (90.00 on GPT-4 train5000) and **UDante** **91.46%** (85.8 on GPT-4 train200). For **LLCT** the best score **99.66%** (94.5 on GPT-4 train1000) is produced by the Trankit individually trained model. For UAS and LAS the results are compared to best numbers reported by [Gamba and Zeman \(2023a\)](#). They have accomplished this using jackknifing technique. In this the training data is divided into n parts, where n-1 parts are used to train a model to annotate the remaining nth part. When rotating this n times, we receive a version of the whole training data with predicted annotations, which can be used during final model training. Therefore, the final model is trained using predicted annotations, in this case the dependency parsing model is trained using predicted morphology and lemmas.

From N-grams to Pre-trained Multilingual Models For Language Identification

Thapelo Sindane¹ and Vukosi Marivate^{1,2}

¹Data Science for Social Impact,
University of Pretoria, South Africa

²Lelapa AI
sindane.thapelo@tuks.co.za,
vukosi.marivate@cs.up.ac.za

Abstract

In this paper, we investigate the use of N-gram models and Large Pre-trained Multilingual models for Language Identification (LID) across 11 South African languages. For N-gram models, this study shows that effective data size selection remains crucial for establishing effective frequency distributions of the target languages, that efficiently model each language, thus, improving language ranking. For pre-trained multilingual models, we conduct extensive experiments covering a diverse set of massively pre-trained multilingual (PLM) models – mBERT, RemBERT, XLM-r, and Afri-centric multilingual models – AfriBERTa, Afro-XLMr, AfroLM, and Serengeti. We further compare these models with available large-scale Language Identification tools: Compact Language Detector v3 (CLD V3), AfroLID, GlotLID, and OpenLID to highlight the importance of focused-based LID. From these, we show that Serengeti is a superior model across models: N-grams to Transformers on average. Moreover, we propose a lightweight BERT-based LID model (za_BERT_lid) trained with NHCLT + Vukzenzele corpus, which performs on par with our best-performing Afri-centric models.

1 Introduction

Automatic language identification (LID) is the task of determining the underlying natural language used in a written or spoken corpus (McNamee, 2005). This is a challenging problem, especially for languages with insufficient training examples and closely related languages, particularly low-resourced languages (Haas and Derczynski, 2021). For South African languages, building quality LID technologies is significantly important for sourcing internet data, which has served as a de-facto repository for many low-resourced languages, especially from public domains such as news websites (Marivate and Sefara, 2020; Adelani et al., 2021; Dione

et al., 2023; Adelani et al., 2023; Lastrucci et al., 2023).

Statistical approaches for automatic LID such as N-grams (Dube and Suleman, 2019), and more classical machine learning models such as Logistic Regression, Naive Bayes, Random Forest, Boosting machines, Support Vector Machines, and Clustering techniques (e.g K Nearest Neighbors) have been proposed (Haas and Derczynski, 2021). Moreover, contemporary neural-based architectures such as deep neural networks and convolutional neural networks have also been tested. In all cases, not enough work for the South African languages is reported.

On the other hand, recent algorithmic advancements such as transformer architectures have made a significant impact on the Natural Language Processing landscape (Devlin et al., 2018; Conneau et al., 2019). With this sudden shift in perspective, many works have proposed automatic LID using large pre-trained multilingual models, derived from attention mechanisms (Vaswani et al., 2017). Large pre-trained multilingual models are transformer-based architectures simultaneously trained on multiple languages (hence multi-lingual) using various techniques such as token (s) masking training technique, where tokens from a given sentence example are hidden and the objective of the training transformer is to predict the hidden word (s).

In this work, we make use of the recently released Vuk'zenzele crawled corpus (Lastrucci et al., 2023) and the NCHLT dataset (Eiselen and Puttkammer, 2014) to develop and experiment on automatic language identification models on 10 low-resourced South African languages: Northern Sesotho (nso), Setswana (tsn), Sesotho (sot), isiZulu (zul), isiXhosa (xho), isiSwati (ssw), isiNdebele (nbl), Tshivenda (ven), Xitsonga (tso), and Afrikaans (af). Additionally, we included the high-resource South African English (eng) to ensure representation of all 11 official languages in South

Africa. We conduct extensive experiments on N-gram models, large pre-trained multilingual models – XLM-r, mBERT, and Afri-centric multilingual models – AfriBERTa, Afro-XLMr, AfroLM, and Serengeti. We shed light on the limitations and robustness of N-gram-based approaches and the significant improvement boost of pre-trained multilingual models, especially, for those pre-exposed to low-resourced South African languages during pre-training.

2 Related Work

Large pre-trained multilingual models have shown astonishing state-of-the-art results on various Natural Language Processing (NLP) tasks such as Machine Translation, Question Answering, and Sentiment Analyses (Stickland et al., 2021; Yang et al., 2019; Adebara et al., 2023b). A precursor of these tasks is the crawling of large volumes of internet data and categorizing the data into different languages (i.e. language identification) for pre-training. For language identification, many works have used pre-trained multilingual models to expand monolingual datasets using the internet.

Jauhiainen et al. (2021) conducted a comparative study between adaptive Naive Bayes, HeLI2.0, multilingual BERT, and XLM-r models for Dravidian language identification in a code-switched context (i.e. a conventional *modus operandi* for communication on the internet). Caswell et al. (2020) developed a transformer-based LID model aside from basic filtering techniques such as tunable-precision-based filtering using a created wordlist, TF-IDF filtering, and a percent-threshold filtering threshold proposed in their study to filter noisy web-crawled content. Although they were able to collect corpora for over 212 languages, their set-up for their best-performing transformer model was unclear. Similar to our work, Kumar et al. (2023) conducted a comparative study on DistilBERT, ALBERT, and XLM-r and showed that a lightweight version of DistilBERT delivers comparable results to resource-intense models. Adebara et al. (2022), on the other hand, implemented a massive transformer-based LID model with 12 attention layers and heads. They then trained this model on 512 languages with close to 2 million sentences across 14 language families (South African languages included). Their model achieved over 95 % F1 score on a left-out test sets, outperforming available LID tools: CLD version 2, Langid, Fast-

Corpora	No. Sent	Voc	Unq. Voc	Train	Dev	Test
Vuk	33K	690K	132K	3395	-	728
NCHLT + Vuk	74K	16M	258K	6790	-	1454

Table 1: Corpora statistics for Vuk and NCHLT

text, etc. Kargaran et al. (2023) created a language identifier that covers a whopping 1600 low-high-resourced African languages. Due to the unavailability of resources utilized in previous studies, our research concentrated exclusively on 11 South African languages, with only 3 language families - Sotho-Tswana, Nguni, and Creole. Furthermore, we will only consider a comparison of diverse pre-trained multilingual models (E.g mBERT, XLMr, AfriBERTa, Afro-XLMr, Serengeti, e.t.c) and two lightweight BERT-based models – DistilBERT, and `za_BERT_lid` model.

3 Methodology

The methodology employed in this study uses language-identifiable monolingual corpora from reliable sources as training examples for language identification and compares various pre-trained multilingual models for the task of discriminating between languages.

3.1 Corpora

Text corpora for the 11 South African languages were acquired from two sources: Vuk’zenzele (Vuk) (Lastrucci et al., 2023) and National Centre for Human Language Technology (NCHLT) corpora (Eiselen and Puttkammer, 2014). Table 1, describes the number of sentences (No. Sent), vocabulary (Voc) sizes, unique vocabulary sizes (Unq. Voc), and the train size per language, development set size, and test size per language splits for corpora Vuk and NCHLT. We ensure consistent train and test examples across all languages, by ensuring that all train, and test examples for each language are equal. Therefore, we only had varying development sizes. Additionally, we only considered sentences in the range of 3-50 tokens and did not use the rest of the corpus. Figure 1, and 2, describe the sentence length distribution for Vuk, and NCHLT corpora respectively.

3.2 Pre-processing

The dataset is observed to contain links, digits and therefore our pre-processing included the removal of URLs, digits, punctuations, and followed by lower-casing all sentences using Python regular

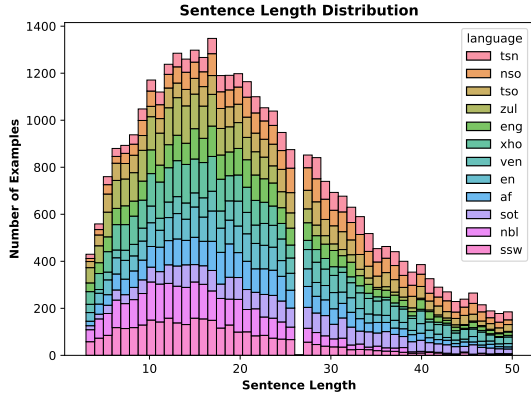


Figure 1: Sentence length distribution of Vuk corpora. The x-axis denotes the number of tokens (words) in the sentences.

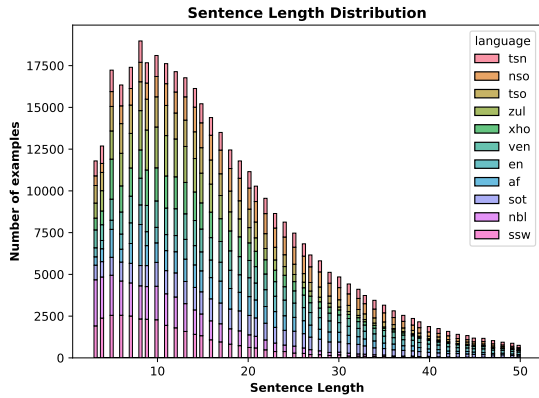


Figure 2: Sentence length distribution of NCHLT + Vuk corpora. The x-axis denotes the number of tokens (words) in the sentences.

expressions. Special characters such as š, found in Northern Sotho were left intact.

3.3 Language detection algorithms

3.3.1 N-grams

An N-gram is a sequence of consecutive characters from text (Dube and Suleman, 2019). This study explored character Bi-grams (2 consecutive characters), Tri-grams (3 consecutive characters), and Quad-grams (4 consecutive characters) models. We build each model for each language from the training dataset (Vuk, NCHLT, and Vuk + NCHLT). Furthermore, we experimented with various data sizes to investigate the impact of the number of training examples on N-gram models and this showed a performance ceiling, where an increase in training examples does not significantly

impact the quality of the models (shown in Figure 3). Each model is made up of a list of tuples of characters-frequency pair ordered in descending order.

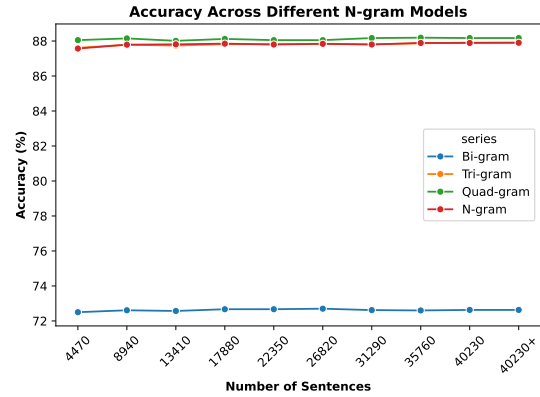


Figure 3: Data size variation performance on Vuk test data.

To discriminate between languages, the models use a ranking function. The ranking function calculates the distance of the frequency distributions of the input examples from the existing N-gram model's frequency distributions (with $k=50$ as the number of ordered N-grams to consider from the trained N-grams). The frequency distribution is calculated as the number of occurrences of each observed N-gram divided by the total number of N-grams from the corpus and taking the log of that ratio. For a given input example (in Northern Sotho) "Ke ya go thopa sefoka" translation - "I am going to win the trophy", the model first extracts the character N-grams (e.g. 2 characters if the observed model is Bi-gram) – Bi-gram Output: ['ke', 'ya', 'a_', 'go', 'th', 'ho', 'op', 'pa', 'se', 'ef', 'fo', 'ok', 'e_', '_y', '_g', 'o_', '_t', '_s'] sorted in reverse, and then the frequency distribution from the existing trained models (looking only at 50 top N-grams per language) for all the languages are compared with the new frequency distribution of the input sentence and the one with the closest similarity is considered the language of the input example. Figure 4, 5, and 6 presents heatmaps depicting the probability scores generated by the ranking function exclusively for all test examples, correctly predicted sentences, and incorrectly predicted examples, during the test phase respectively. The heatmaps reveal that the concentration of scores ranges between 0.04 and 0.06, which could be further used to drive a model's outcome improving the confidence in predictions. This

observation suggests that ranking functions play a crucial role in N-gram-based models, warranting further investigation.

N-grams experimental setup We experimented with Bi (2), Tri (3), and Quad (4) consecutive character sequences to build our models. Additionally, we combined all 3 and called it N-grams combined.

3.3.2 Naive Bayes Classifier

Naive Bayes have been the default standard for various LID tasks such as code-switching detection, dialect discrimination, word-level language detection, and e.t.c. (Dube and Suleman, 2019; Jauhainen et al., 2019). In this study, we experimented with the multinomial Naive Bayes Classifier (NBC) implementation from Python’s scikit-learn. With NBC, we were able to extract discriminating features per language, supporting model prediction (Figure 8), and significantly improved on N-gram models (see confusion matrix in Figure 9). This highlighted important feature correlation, especially for related languages, which explains why it is challenging to discriminate among closely related languages. Moreover, this highlights the importance of lexicon-driven approaches for language filtering mentioned in Caswell et al. (2020) as alternative measures to mitigate these ambiguities.

Naive Bayes Classifier experimental setup We experimented with a TF-IDF vectorizer to generate input features. For this, we used the character bi-gram, tri-gram, quad-gram, and the 3 types combined as consecutive subwords to generate TF-IDF features. We also generated word level input features using CountVectorizer. We used a multinomial version of the Naive Bayes classifier with mostly default parameters from scikit-learn (except the alpha parameter where we tested $\alpha = 0.0001, 1.0$, where $\alpha = 1.0$ performed better). Finally, we trained Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Logistic Regression with the same input features and their scikit-learn default parameters to compare performance outcomes with NBC.

3.3.3 Pre-trained Multilingual Models

This study explored a diverse set of massively pre-trained multilingual models: mBERT, XLM-r, RemBERT, and their Afri-centric counterparts: AfriBERTa, Afro-XLMr, AfroLM, and Serengeti due to their enhanced text processing capabilities and their ability to handle low-resourced languages

with complex linguistic nuances (Devlin et al., 2018; Conneau et al., 2019; Ogueji et al., 2021; Alabi et al., 2022; Dossou et al., 2022; Adebara et al., 2023a).

Large pre-trained multilingual models experimental setup Following setups in (Adelani et al., 2023; Dione et al., 2023), we used a batch size of 16, a learning rate of $2e^{-5}$, 20 epochs, save step of 10000, and sequences cut-off of 200 for all models. We ran our experiments five times with different seeds { 1,.., 5 } and reported the average results.

4 Results

4.1 Baselines

Table 2, shows results for baseline models Bi-gram, Tri-gram, Quad-gram, N-gram combined (N-gram Comb) – which uses bi-, tri-, and quad- -grams combined, and Naive Bayes Classifier (NBC) with the same character N-grams. Naive Bayes with word-level features outperform the rest of the baseline models. Interestingly, for NBC, increasing the character spans improves the performance of the classifier. Figure 10, 11, 12, 13, 14, and 15 depicts the impact of increasing the data size on models NBC, Support Vector Machine (SVM), and Logistic Regression (Log Reg) on various training features – uni-grams, bi-grams, tri-grams, quad-grams, N-grams combined, and word-level features derived using TF-IDF respectively. NBC, SVM, and Log Reg show improved performance with the change in input features while the training size shows gradual improvement in accuracy. KNN was also tried, however, the model showed abysmal performance across all features except for Bi-gram input features and was therefore omitted from the plots.

In the N-gram class, the Quad-gram ranking outperforms the rest of the N-gram-based models. Figure 17, depicts the impact of sentence length on N-gram models performance. This shows that the group of N-gram models struggles to classify shorter sentences, while NBC performs slightly better with them (Figure 18). This may be due to shorter sentences not carrying enough signal information for N-grams to discriminate across all languages as mentioned in Haas and Derczynski (2021). Additionally, N-gram-based models depict inconsistent performance across languages, where improved performance is achieved for select languages and for a specific N-gram type (E.g Bigram – eng, ven, af, e.t.c, Tri-gram – eng, tso, nso, e.t.c),

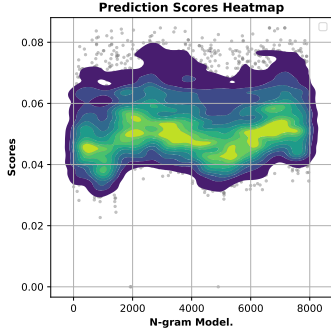


Figure 4: Score heatmap for all predictions using N-gram

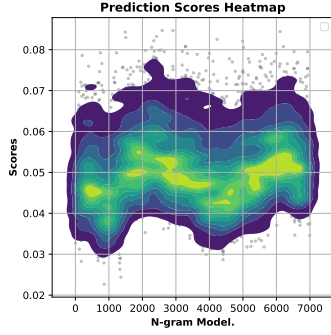


Figure 5: Score heatmap for correctly predicted examples using N-gram

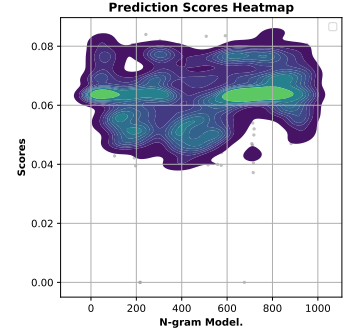


Figure 6: Score heatmap for incorrectly predicted examples using N-gram

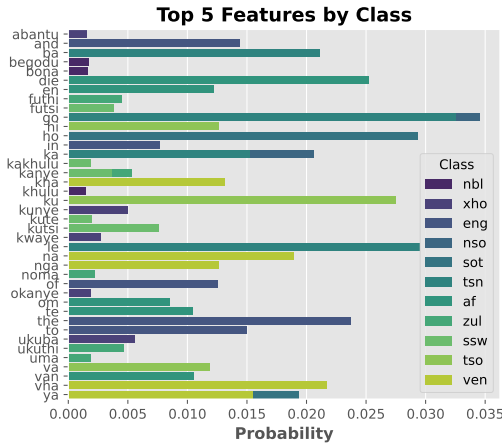


Figure 8: Top 5 important features per class from Naive Bayes

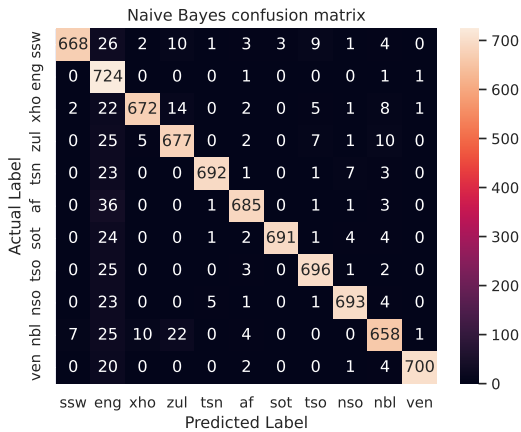


Figure 9: Accuracy of Naive Bayes Classifier.

while other languages underperform (e.g zul, isiNdebele (nbl)) (see Figure 19). Furthermore, the complexity of LID is exacerbated by closely related

Baseline	Acc	Prec	Rec	F1
Vuk				
Bi-gram	72.7	73.5	72.6	72.3
Tri-gram	87.9	88.4	87.9	88.1
Quad-gram	88.4	88.9	88.4	88.5
N-gram (Comb)	87.8	88.3	87.8	88.0
NBC (word-level)	94.5	95.2	94.5	94.6
NBC (2)	90.2	90.7	90.2	90.4
NBC (3)	93.4	93.8	93.4	93.5
NBC (4)	94.4	94.8	94.4	94.5
NBC (Comb)	94.0	94.5	94.0	94.1
K NN (2)	85.0	85.0	85.0	85.0
Log Reg (4)	94.0	95.0	94.0	94.0
SVM (4 & 2-4)	94.0	95.0	94.0	94.0

Table 2: Baseline performance evaluation using Accuracy (Acc), F1 score (F1), Precision (Prec), and Recall (Rec). K Nearest Neighbor (K NN), Logistic Regression (LR), and Support Vector Machine (SVM) are reported with best feature inputs bi-gram (2), quad-grams (4), and combinations (2-4) respectively.

languages (see confusion matrix in Figs. 20–23). While varying dataset size, and character N-gram choices slightly improve performance on distinguishing among closely related languages (Figure 3), it does not add any significant improvement on a per-language basis (see Figure 19), where languages such as isiZulu (zul) are showing no further improvement. For this, we explore large pre-trained multilingual models for automatic LID in the next subsection.

4.2 Pre-trained Multilingual Models

Table 3 reports the accuracy (Acc), precision (Prec), recall (Rec), and F1 score (F1) of pre-trained multilingual models: mBERT, XLM-r, RemBERT;

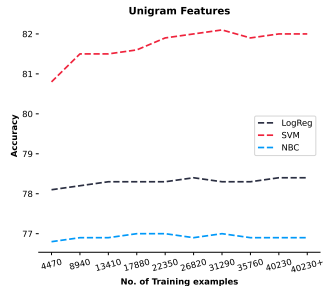


Figure 10: Unigram

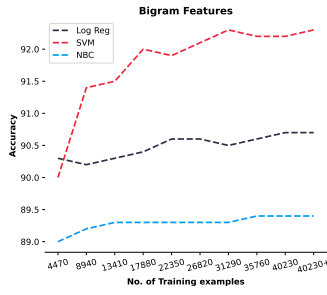


Figure 11: Bi-gram

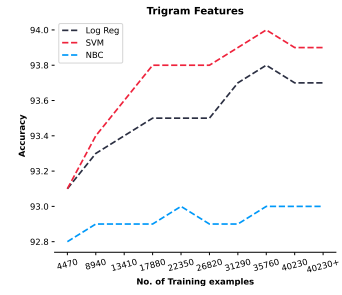


Figure 12: Tri-gram

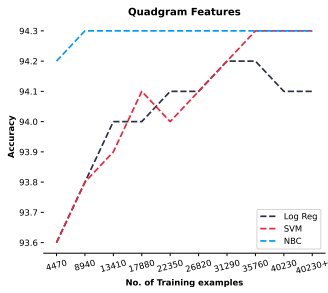


Figure 13: Quad-gram

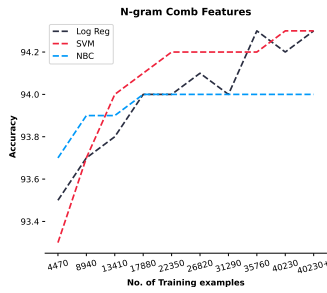


Figure 14: N-grams Comb

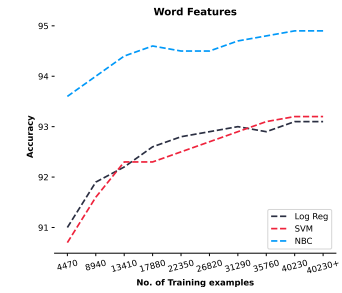


Figure 15: Word-level

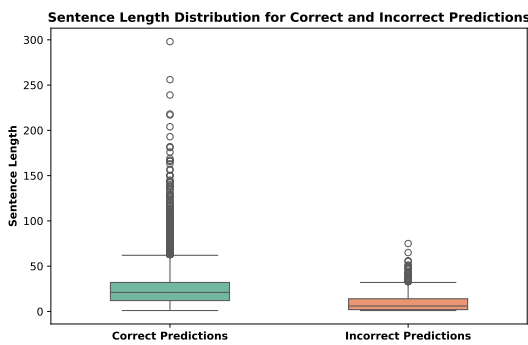


Figure 17: Box diagram depicting sentence length of correctly predicted and incorrectly predicted sentences.

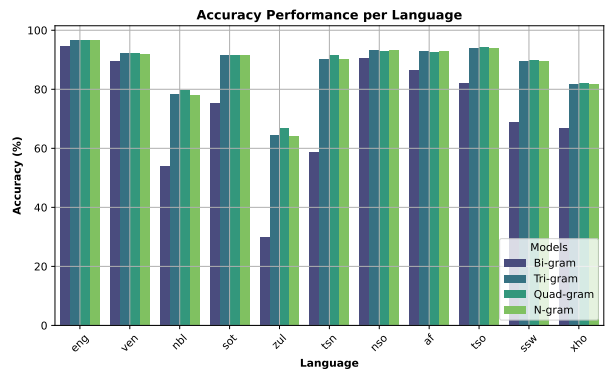


Figure 19: Accuracy score per language using N-grams

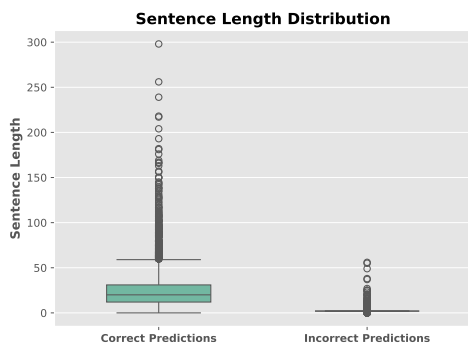


Figure 18: Incorrectly and correctly NBC classified sentence lengths

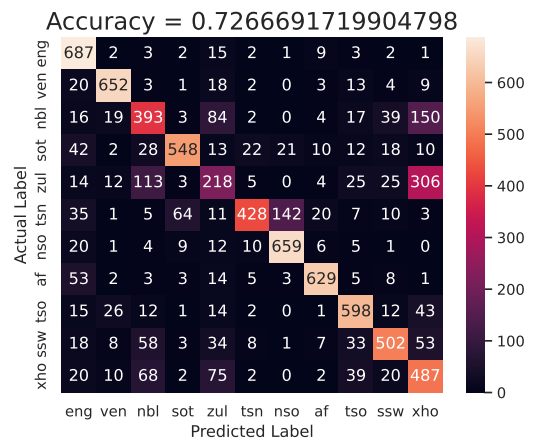


Figure 20: Bi-gram Confusion matrix on Vuk test data

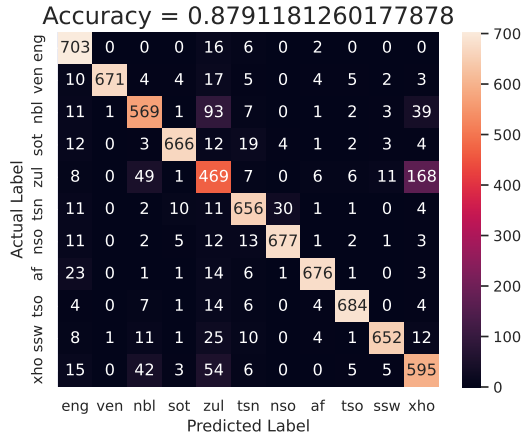


Figure 21: Tri-gram Confusion matrix on Vuk test data

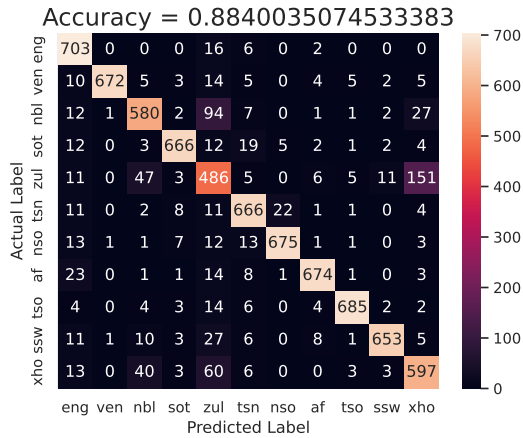


Figure 22: Quad-gram Confusion matrix on Vuk test data

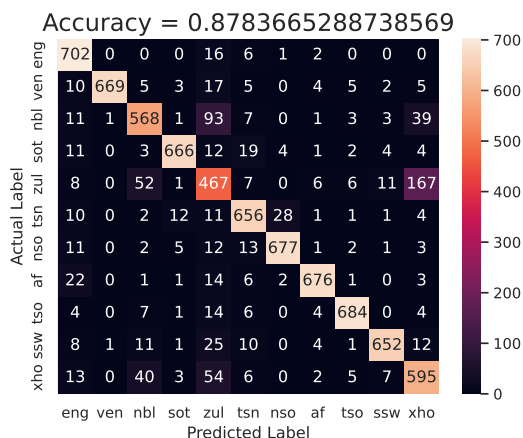


Figure 23: Accuracy of N-gram type (Bi-gram, Tri-gram, Quad-gram) combined

Afri-centric pre-trained models: AfriBERTa, AfroXLMr, AfroLM, and Serengeti; publicly available

LID tools covering South African languages: Compact Language Detector (CLD) version 3 (V3), AfroLID (Adebara et al., 2022), GlotLID (Kargaran et al., 2023), and OpenLID (Burchell et al., 2023); and our proposed lightweight BERT-based architectures: za-BERT-lid, and DistilBERT.

Pre-trained-multilingual models show impressive results for this task, with over 90% average accuracy. Serengeti outperforms the rest of the models with an average accuracy of 98 %, while mBERT is the least-performing model with an average accuracy of 96 % (≈ 2 points difference). Most importantly, the group of Afri-centric models outperforms the largely pre-trained multilingual models with the best model (XLMr-large) in this category performing slightly worse than the lowest performing model (AfroLM) in the Afri-centric group. Moreover, our proposed za-BERT-lid, and DistilBERT perform on par with the best-performing model (≈ 2 points difference) despite them being much smaller in size.

On the other hand, available LID tools show impressive and incremental results. For these models, GlotLID outperforms the rest of the sampled models in this study. This may be due to GlotLID being trained on Vuk data, giving the model an unfair advantage over others. Despite this, analyses of the predictions show that the compared models are not completely wrong, as they often struggle with closely related languages such as Sotho-Tswana language family {nso, sot, tsn}, and Nguni languages {xho, zul, ssw, and nbl}. Perhaps to remedy this, the training of LID models should prioritize precision as a metric of evaluation. Noticeably, but not alarming, the LID tools also predict unrelated languages from their training list, which perhaps highlights the need for a more focused approach rather than including many languages at once. However, we feel this claim needs further justification and we will consider this in future work.

4.3 Cross-domain evaluation

We also wanted to test our model on cross-domain datasets to inspect their generalization capabilities. We simulated this by training with Vuk data and tested it on NCHLT, and vice versa. Table 4 reports the performance of pre-trained models for examining the cross-domain evaluation theory. This table shows that the performance of the multilingual models trained with Vuk and tested with NCHLT dropped by approximately (4%-5%) across all mod-

Model	Acc	Prec	Rec	F1
PLM				
mBERT	96.7	96.7	96.6	96.7
XLMr-base	97.1	97.1	97.1	97.1
XLMr-large	97.3	97.3	97.3	97.3
RemBERT	97.1	97.1	97.1	97.1
Afri-centric				
AfriBERTa	97.6	97.6	97.6	97.6
Afro-XLMr-base	97.7	97.8	97.7	97.7
Afro-XLMr-large	98.0	98.0	98.0	98.0
AfroLM	97.4	97.5	97.4	97.4
Serengeti	98.3	98.3	98.3	98.3
LID Tools				
CLD V3	40.2	33.6	40.2	35.7
AfroLID	66.1	72.1	66.1	64.2
OpenLID	80.8	71.7	80.8	75.0
GlottLID	97.5	98.3	97.5	97.9
Lightweight				
za-BERT-lid	96.8	96.8	96.8	96.8
DistilBERT	96.2	96.2	96.2	96.2

Table 3: Performance evaluation scores of pre-trained multilingual models, available LID tools, and lightweight BERT-based models averaged over 5 runs per metric.

els. In contrast, training with NCHLT and testing with Vuk showed performance improvements. This could be due to NCHLT having more training examples, and a large vocabulary (see Table 1) allowing the model to learn more nuanced representations. Notably, larger models show better performance over smaller models for this task.

5 Discussions

Ensuring the development of robust LID detection systems remains a critical research area with implications on many NLP tasks. Importantly, the availability of reliable LID systems ensures accurate reporting on the state of low-resourced languages (Kreutzer et al., 2022).

On the side of model performance, baseline techniques such as Naive Bayes, Support vector Machines, and Logistic Regression seem to be performing quite well on the task of sentence-level language identification. We recommend these models for further research for high-level LID, compared to large pre-trained multi-lingual models which require specialized computing resources such as GPUs, to accelerate training. However, we deem such trade-offs to require more research, especially

Model	Vuk Test	NCHLT Test
Vuk Trained		
mBERT	-	91.0
XLMr-base	-	91.4
XLMr-large	-	92.2
RemBERT	-	92.3
AfriBERTa	-	92.1
Afro-XLMr-base	-	93.6
Afro-XLMr-large	-	94.1
AfroLM	-	91.8
Serengeti	-	94.9
za-BERT-lid	-	91.3
DistilBERT	-	90.9
NCHLT Trained		
XLMr-base	95.6	93.2
Afro-XLMr-base	96.3	93.6
Serengeti	97.7	94.8

Table 4: Cross-domain evaluation of models trained with Vuk and tested with NCHLT and vice-versa. Reported in F1 score averaged over five runs

in complex LID subtasks such as code-switching, or similar language discrimination.

We also, highlight the importance of evaluation metric selection as we have observed that most of the LID tools explored in this study are not completely wrong, but rather have challenges discriminating among closely related languages. Therefore, we recommend precision as an evaluation metric for LID to be further investigated.

6 Conclusion

Language Identification remains a critical study area for the widespread inclusion of many low-resourced languages into the booming technology space. In this study, we experimented with statistical approaches, traditional machine learning techniques, the recent advanced pre-trained multilingual models, as well as LID tools publicly available (covering a wide range of African languages) on the task of LID for 11 South African language discrimination. We were able to shed light on the approaches showing promising results in the South African language context and made suggestions for future directions. Concretely, we showed that the Naive Bayes algorithm performs surprisingly well for LID and warrants further exploration and exploitation, especially given its cheap-compute advantage. Finally, we compared publicly available pre-trained models and showed that context-

exposed models have an edge over other context-oblivious multilingual models, where context refers to the language. We released our models on [HuggingFace](#) and code with datasets on [GitHub](#).

7 Limitations

In this study, we did not explore any use of word embeddings for language identification. Word embeddings played in crucial role in the development of language technologies, and it would have been interesting to experiment with them. However, such resources are not readily available for many low-resourced languages.

Aside from experimenting and getting results for other traditional models such as Logistic regression, K Nearest Neighbor, and Support Vector Machines, it would have been interesting to develop and experiment with deep neural networks such as multi-layered perceptions, and convolutional neural networks. As universal approximators, these models tend to produce desirable results, with the caveat of requiring time for hyper-parameter tuning.

This study did not extensively explore the impact k (used 50 for this study), which is the count of the N-grams list used to calculate the ranking. However, we aim to explore this extensively in future works.

It is known that LID techniques tend to overfit to domain data, and therefore it would have been interesting to create free-text data created by humans and test the generalization capabilities of the developed models on human-generated text.

Recent studies have focused on resource-conscience alternatives for either compute efficiency, parameter reduction, etc. It would have been interesting if this work would have explored the recently active approaches focusing on smaller models utilizing parameter transfer, and adaptations (Kumar et al., 2023). However, these techniques require intense hyper-parameter selection and tuning, and slightly longer training times, which was not in the scope of this study.

Finally, we aim to incorporate BANTUBERT¹, and zaBANTUBERT² models trained with monolingual South African corpora in our future work.

¹<https://huggingface.co/dsfsi/BantuBERTa>

²<https://huggingface.co/dsfsi/zabantu-xlm-roberta>

Acknowledgements

The authors would like to thank the ABSA chair of Data Science for funding and support. We also like to express our gratitude to Matimba Shingange and Michelle Terblanche for their input and reviews of the paper. Finally, the authors would like to thank the Data Science for Social Impact Research lab for their continued support.

References

- Ife Adebbara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023a. [SERENGETI: Massively multilingual language models for Africa](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.
- Ife Adebbara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*.
- Ife Adebbara, Abdelrahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023b. [Serengeti: Massively multilingual language models for africa](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537.
- David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. 2023. [Masakhanews: News topic classification for african languages](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 144–159. Association for Computational Linguistics (ACL).
- David Ifeoluwa Adelani, Dana Ruiters, Jesujoba O Alabi, Damilola Adebajo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021. [Menyo-20k: A multi-domain english-yorùbá corpus for machine translation and domain adaptation](#). In *AfricaNLP*.
- Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. [Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An open dataset and model for language identification](#). In *Proceedings of the 61st Annual Meeting of the Association for*

- Computational Linguistics (Volume 2: Short Papers)*, pages 865–879.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Cheikh M Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, et al. 2023. Masakhapos: Part-of-speech tagging for typologically diverse african languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900.
- Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustainLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Meluleki Dube and Hussein Suleman. 2019. Language identification for south african bantu languages using rank order statistics. In *Digital Libraries at the Crossroads of Digital Information for the Future: 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4–7, 2019, Proceedings 21*, pages 283–289. Springer.
- Roald Eiselen and Martin J Puttkammer. 2014. Developing text resources for ten south african languages. In *LREC*, pages 3698–3703. Citeseer.
- René Haas and Leon Derczynski. 2021. Discriminating between similar nordic languages. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 67–75. Association for Computational Linguistics.
- Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019. Discriminating between mandarin chinese and swiss-german varieties using adaptive language models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187.
- Tommi Jauhiainen, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Comparing approaches to dravidian language identification. *arXiv preprint arXiv:2103.05552*.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Sch"utze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.
- R Prasanna Kumar, R Elakkiya, R Venkatakrisnan, Harrieni Shankar, Y Sree Harshitha, K Harini, M Nikhil Reddy, et al. 2023. Transformer-based models for language identification: A comparative study. In *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–6. IEEE.
- Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. 2023. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 18–25.
- Vukosi Marivate and Tshephisho Sefara. 2020. Improving short text classification through global augmentation methods. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 385–399. Springer.
- Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of computing sciences in colleges*, 20(3):94–101.
- Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *NAACL HLT 2019*, page 72.

Visualising Changes in Semantic Neighbourhoods of English Noun Compounds over Time

Malak Rassem, Myrto Tsigkouli, Chris Jenkins, Filip Miletić, Sabine Schulte im Walde

Institute for Natural Language Processing, University of Stuttgart, Germany

{malak.rassem, myrto.tsigkouli}@outlook.com

{christopher.jenkins, filip.miletic, schulte}@ims.uni-stuttgart.de

Abstract

This paper provides a framework and tool set for computing and visualising dynamic, time-specific semantic neighbourhoods of English noun-noun compounds and their constituents over time. Our framework not only identifies salient vector-space dimensions and neighbours in notoriously sparse data: we specifically bring together changes in meaning aspects and degrees of (non-)compositionality.

1 Introduction

Noun compounds (NCs) such as *fairy tale* and *gold mine* represent a type of multiword expressions (MWEs) whose meanings are semantically idiosyncratic to some degree, i.e., their meanings are not necessarily fully predictable from the meanings of their parts (Partee, 1984; Sag et al., 2002; Baldwin and Kim, 2010). While the restricted compositionality of NCs has been explored extensively and across research disciplines from synchronic perspectives (Mitchell and Lapata, 2010; Reddy et al., 2011; Schulte im Walde et al., 2013, 2016; Baroni et al., 2014; Cordeiro et al., 2019; Garcia et al., 2021; Miletić and Schulte im Walde, 2023, i.a.), the field is still lacking an adequate amount of empirical large-scale approaches towards diachronic models, in order to explore the emergence and changes of compound meanings over time. Up to date, Dhar et al. (2019) and Dhar and van der Plas (2019) exploited the Google n -gram corpus and information-theoretic as well as cosine distance measures to predict the compositionality of the compounds in Reddy et al. (2011), and to detect novel compounds, respectively. Maurer et al. (2023) investigated whether diachronic developments of the frequencies and productivities of the compounds and their constituents in Cordeiro et al. (2019) are salient indicators of the NCs' present-day degrees of compositionality, and Mahdizadeh Sani et al. (2024) applied standard cosine-based

measures of (dis)similarity to the same compounds and constituents over time.

The current study contributes to the so far limited computational models and insights on the diachronic development of NC compositionality. We provide a framework and a tool set for computing and visualising semantic neighbourhoods of English NCs over time. By focusing on semantic neighbours we explicitly target changes in the salient meaning aspects of NCs; more specifically, we bring together semantic neighbourhoods of NCs and their constituents, thus aligning changes in meaning aspects and degrees of compound-constituent (non-)compositionality. A major side-challenge is concerned with identifying an appropriate set of vector-space dimensions, both regarding the semantic interpretations of the dimensions and regarding the notorious sparse-data problem in historical corpus data that strongly affects compound representations. The contributions of this paper are the following.

- **Semantic Space:** A carefully crafted semantic vector space to represent those 195 noun-noun compounds and their constituents from Cordeiro et al. (2019) that occurred in all time slices of the cleaned corpus of historical American English CCOHA (Davies, 2012; Alatrash et al., 2020).
- **Semantic Neighbours:** Semantic neighbourhoods for compounds and their constituents, both (i) time-specific and dynamic as well as (ii) static present-day representations.
- **Temporal Compound-Constituent Visualisation Tool:** An adaptation of a deterministic approach to multi-dimensional scaling and two-dimensional plotting (Hilpert, 2016; Tsigkouli, 2021) to the vector-space representations of compounds, constituents and semantic neighbourhoods.

The semantic spaces and neighbours of our English compounds and constituents, as well as the visualisation tool which is applicable to also further compound and constituent targets in English and additional languages, are publicly available from <https://www.ims.uni-stuttgart.de/data/dia-neighbour-nn>.

2 Data

2.1 Corpus: CCOHA

As our diachronic text corpus resource, we used the clean version of the Corpus of Historical American English, referred to as CCOHA (Davies, 2012; Altrash et al., 2020), in order to ensure that the dataset is free from inconsistent lemmas, malformed tokens and other anomalies that could potentially affect the analyses. We then reduced the fine-grained part-of-speech tags in CCOHA to a coarser-grain set of tags, for example, collapsing all variants of nouns like singular common noun (NN1), plural common noun (NN2), singular locative noun (NNL1), etc. under a single broad noun tag NN to generalise the tokens' part-of-speech (POS) tags. A full list of the mapping of the POS tags can be found in Appendix A. To analyse changes over time, the data was segmented into specific timeslices. The selected timeslices are: 1810–1829, 1830–1859, 1860–1889, 1890–1919, 1920–1949, 1950–1979, and 1980–2009, with each range being inclusive.

2.2 Noun Compound (NC) Targets

Our goal is to investigate the semantic evolution of noun compounds across different historical periods, focusing specifically on the noun-noun compounds identified by Cordeiro et al. (2019). Out of the 210 noun-noun¹ compounds mentioned in their work, 195 are present in our corpus. We consider both space-separated and dash-separated compounds, treating equivalents like *credit card* and *credit-card* as identical entities for our analytical purposes. In order to exclude compounds with more than two constituents, we imposed a restriction on the POS tag patterns; namely, the tokens immediately preceding and succeeding a noun-noun target compound must not be tagged as nouns (NN) for the sequence to qualify as a two-part noun-noun compound.

¹We disregarded noun compounds with other than nominal modifiers (such as adjective-noun compounds).

3 Semantic Space and Neighbours

3.1 Semantic Vector-Space Creation

As the backbone of our semantic space for plotting compounds as well as their constituents and semantic neighbours over time, we identified a set of semantic space points (SSPs). These SSPs were defined as nouns appearing with a frequency >500 in the entirety of the CCOHA, i.e., not just within individual timeslices. The threshold was set to ensure a substantial enough occurrence for meaningful semantic analysis. Then the top 50 most frequent nouns were excluded from the SSPs to eliminate potential semantic hubs (Radovanović et al., 2010; Dinu et al., 2015) that could dominate the analysis due to their high rate of occurrence, given that they typically represent semantically generic terms. Our criteria resulted in identifying 9,345 unique nouns that served as SSPs for further analysis.

For all noun compounds, their constituents and all SSPs, we computed timeslice-specific co-occurrences (TSCs) within a ± 10 -word window. These TSCs were further refined by limiting the context words to those tagged with the reduced POS content tags: nouns (NN), verbs (VV), adverbs (RR), and adjectives (JJ). The TSCs were then transformed into vectorised formats to enable further processing. This conversion entails mapping the co-occurrence data into numerical vectors, with each dimension corresponding to a specific context word. The magnitude in each dimension was determined by the frequency of each context word's co-occurrence with the noun compounds, constituents or SSPs within the defined timeslice.

We chose to use simple frequency counts for co-occurrences rather than alternative association measures (Evert, 2005) due to the complexities and potential mathematical incorrectness involved. Specifically, measures such as variants of mutual information would require division by the total number of all co-occurrences of the targets we are dealing with. In our case, this would mean coalescing the noun compounds, their constituents and the SSPs together. However, doing so would lead to double counting, because constituents may also function as SSPs. Moreover, there is considerable overlap between the co-occurrences of compounds and those of their constituents or SSPs. For instance, the co-occurrences for the compound *wedding day* are essentially the identical subset of its constituents, which are also SSPs, thereby leading to redundancy in our counts.

Target	Timeslice	5 Nearest Neighbours
credit card	1830–1850	—
	1920–1940	rationing, gallon, shuttle, questionnaire, invitation
	1980–2000	reservation, card, cash, credit, check
credit	1830–1850	exchange, money, bank, account, circulation
	1920–1940	loan, bank, account, banker, reserve
	1980–2000	card, visa, account, cash, greeting
card	1830–1850	game, paper, trick, minute, stranger
	1920–1940	paper, game, ball, box, trick
	1980–2000	check, credit, paper, line, trick

Table 1: The five nearest neighbours of the compound *credit card* and its constituents, across timeslices.

3.2 Semantic Neighbourhoods

Using cosine (dis)similarity, we compared the TSC vector representations of the noun compounds to those of the SSPs within the same timeslice, in order to quantify their semantic proximity. For each time-specific compound, the five most similar neighbours from the pool of SSPs were identified based on the cosine scores. For example, we can see in Table 1 that the compound *credit card* did not appear in the corpora in earlier timeslices, suggesting that there was no established sense for the compound at that time. Subsequently, the neighbours of *credit card* include written documents, reflecting the term’s initial use to denote means of payment such as traveller’s cheques.² The neighbours in more recent periods transition to the modern sense associated with *cash*.

Following Hamilton et al. (2016), we used as a static semantic space the TSC vectors of the last timeslice of these neighbours, and did the same for those of the compounds’ constituents. This approach allows us to capture the evolving relationships between words over time while maintaining temporally fixed reference points for comparison.

4 Temporal Compound Visualisation

We implemented two methods to visualise time-specific compounds in semantic space.

4.1 Own-Vector Approach

In the own-vector method, we created a single matrix for each compound using its TSC vectors at every timeslice and the TSC vectors of its constituents and neighbours only from the last "static" timeslice. We then applied metric multidimensional scaling

(MDS) to this matrix, which we preferred over non-deterministic approaches such as t-SNE due to its determinism, and derived two-dimensional vector representations for plotting, as previously done by Hilpert (2016) and Tsigkouli (2021). Although the own-vector approach seems to be the most intuitive, we found it to produce objectively sub-optimal plots, where the compounds tend to cluster together and away from the SSPs regardless of the timeslice.

4.2 Projected-Compound Approach

In this refined method, a single matrix is created using the latest (static) timeslice TSC vectors of all compound neighbours and constituents, but excluding the compound’s own TSC vectors. As in the own-vector approach, we derived the coordinates of the neighbours and constituents by applying MDS to this matrix. For the compound, however, rather than using the compound’s own TSC vectors to determine the time-specific coordinates, these vectors were computed as the weighted averages of the respective five time-specific nearest neighbours’ coordinates, with the weights being their cosine scores. The intuition behind this approach was that in the own-vector approach the SSPs’ TSC vectors and the compounds’ TSC vectors consistently clustered away from each other and could not efficiently be visualised together, which we attribute to the severe sparsity in the compound vector representations. In contrast, our refined approach projects a compound’s semantic change over time by reflecting its relative positions to its neighbours’ semantic fields, thus improving over the sparsity issue. Consequently, the method produces plots that more distinctly illustrate the temporal semantic shifts of noun compounds. For example, the trend regarding *credit card* and its neighbours that we described

²<https://www.etymonline.com/word/credit-card>

above based on Table 1 is rather clear in the plot in Figure 1. Likewise, in Figure 2 we observe *gold mine* starting from its literal compositional sense in the earlier timeslices (i.e., the actual mine), where it's surrounded by its constituents, and in later times moving towards SSPs such as *money* and *business*. This shift highlights the development of an additional metaphorical sense of *gold mine* in the later timeslices, as a symbol of value.

5 Conclusion

This study used a corpus-based computational approach to examine the semantic evolution of noun compounds in historical American English, thus contributing to the field of diachronic computational linguistics by providing a methodologically robust tool set for analyzing temporal changes in compound semantics. Future research could expand upon this foundation by exploring other types of multiword expressions.

Limitations

We presented experiments on visualising the temporal evolution of noun compound meanings as captured by high-dimensional semantic vectors. The obtained results strongly depend on the choice of vector space representations and dimensionality reduction methods. We opted for interpretable and deterministic approaches given our linguistic motivation, and with this constraint we explored different implementation variants and presented the most robust systems. Some other combination of experimental settings – including non-deterministic methods – may improve on our results.

More generally, our vector space representations are directly dependent on the properties of the underlying corpus, which is additionally affected by sparsity issues (like most diachronic datasets). A different set of texts may capture different aspects of the target words' semantics; a larger corpus may yield more robust vector representations. Moreover, our experiments are limited to the American English data at our disposal. Due to typological differences in the linguistic realisation of multiword expressions such as noun compounds, our method may not produce equivalent results for other languages or language varieties.

Ethical Considerations

We do not believe that this paper raises ethical issues. We conducted a linguistic analysis of empir-

ically attested data using well-established methods to computationally represent word meaning. Note though that our bottom-up approach automatically induces the semantic neighbours for a specified target word. We therefore cannot exclude the possibility of inadvertently outputting offensive content or depicting societal biases captured by our corpus, which covers American English usage over the course of two centuries. However, we did not encounter these issues in closely inspected results; we also note that they are inherent in any large-scale corpus analysis.

Acknowledgments

The research presented here was supported by the DFG Research Grant SCHU 2580/5 (*Computational Models of the Emergence and Diachronic Change of Multi-Word Expression Meanings*). We also thank Martin Hilpert for providing the starting point for the MDS/R visualisation.

References

- Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean Corpus of Historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.
- Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Boca Raton, USA.
- Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in Space: A Program for Compositional Distributional Semantics. *Linguistic Issues in Language Technologies*, 9(6):5–110.
- Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.
- Mark Davies. 2012. Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English. *Corpora*, 7(2):121–157.
- Prajit Dhar, Janis Pagel, and Lonneke van der Plas. 2019. [Measuring the Compositionality of Noun-Noun Compounds over Time](#). In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 234–239, Florence, Italy. Association for Computational Linguistics.

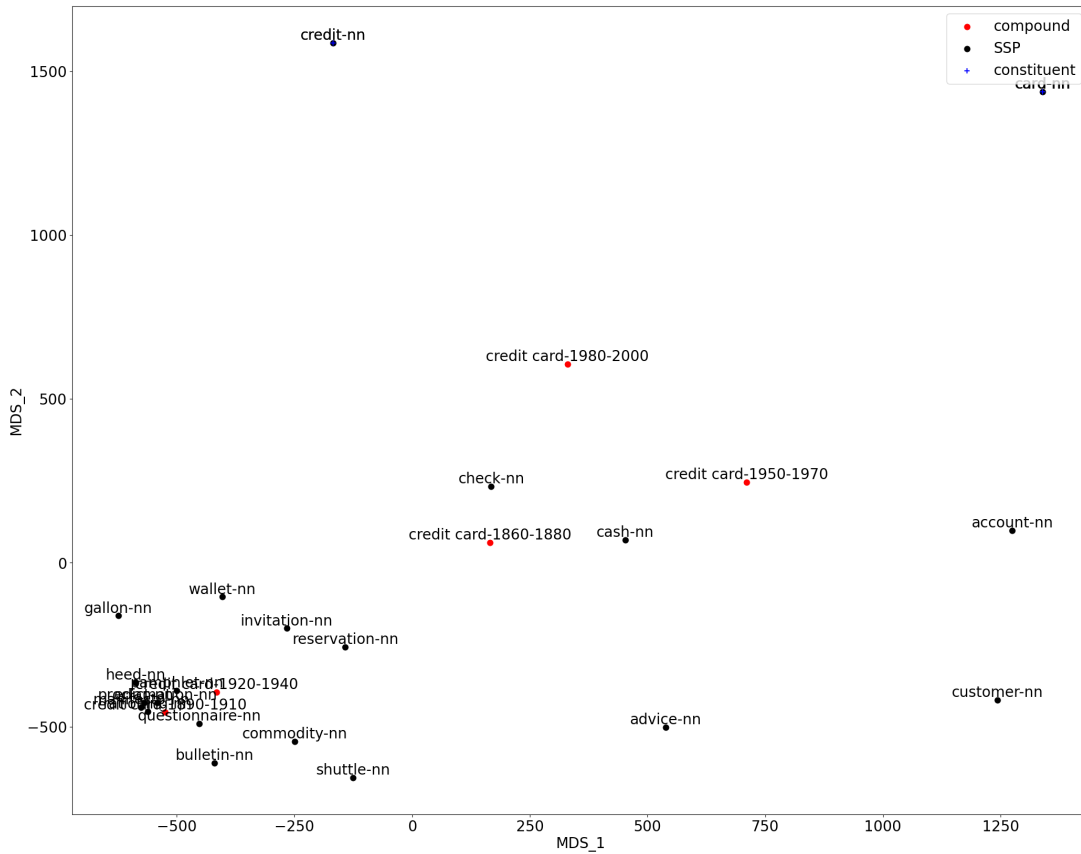


Figure 1: Example of the projected-compound approach on the compound *credit card*.

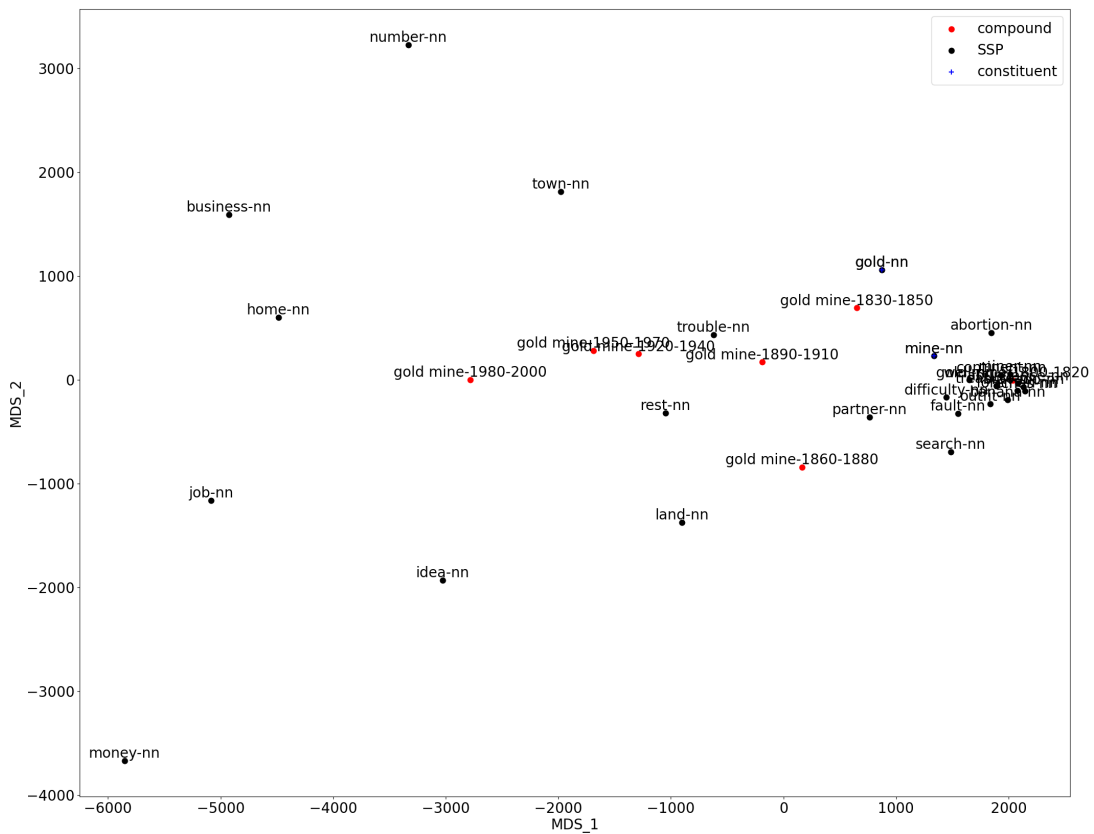


Figure 2: Example of the projected-compound approach on the compound *gold mine*.

- Prajit Dhar and Lonneke van der Plas. 2019. [Learning to Predict Novel Noun-Noun Compounds](#). In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet*, pages 30–39, Florence, Italy. Association for Computational Linguistics.
- Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving Zero-Shot Learning by Mitigating the Hubness Problem. In *Proceedings of the International Conference on Learning Representations, Workshop Track*, San Diego, CA, USA.
- Stefan Evert. 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.
- Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the Representations of Idiomaticity in Vector Models with a Noun Compound Dataset Labeled at Type and Token Levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2730–2741, Online.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Martin Hilpert. 2016. Change in Modal Meanings: Another Look at the Shifting Collocates of *may*. *Constructions and Frames*, 8(1):66–85.
- Samin Mahdizadeh Sani, Malak Rassem, Chris Jenkins, Filip Miletic, and Sabine Schulte im Walde. 2024. What Can Diachronic Contexts and Topics Tell Us About the Present-Day Compositionality of English Noun Compounds? In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 17449–17458, Torino, Italy. European Language Resources Association and International Committee on Computational Linguistics.
- Maximilian Maurer, Chris Jenkins, Filip Miletic, and Sabine Schulte im Walde. 2023. Classifying Noun Compounds for Present-Day Compositionality: Contributions of Diachronic Frequency and Productivity Patterns. In *Proceedings of the 19th Conference on Natural Language Processing*, pages 40–51, Ingolstadt, Germany.
- Filip Miletic and Sabine Schulte im Walde. 2023. A Systematic Search for Compound Semantics in Pre-trained BERT Architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512, Dubrovnik, Croatia. Association for Computational Linguistics.
- Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429.
- Barbara H. Partee. 1984. Compositionality. In Fred Landman and Frank Veltman, editors, *Varieties of Formal Semantics: Proceedings of the 4th Amsterdam Colloquium*, pages 281–311. Foris Publications.
- Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, 11:2487–2531.
- Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Association for Computational Linguistics.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.
- Sabine Schulte im Walde, Anna Hättig, and Stefan Bott. 2016. The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.
- Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. 2013. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA, USA. Association for Computational Linguistics.
- Myrto Tsigkouli. 2021. Studying the Diachronic Changes of Constituent Collocates and Meanings in English Noun Compounds. Master’s thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

A Part-of-Speech Tag Reduction

Table 2 presents the mapping of the original part-of-speech tags into a reduced, coarser set of tags.

Tag Description	Original Tag	Reduced Tag
singular noun of direction (e.g. <i>north, southeast</i>)	ND1	NN
common noun, neutral for number (e.g. <i>sheep, cod, headquarters</i>)	NN	NN
singular common noun (e.g. <i>book, girl</i>)	NN1	NN
plural common noun (e.g. <i>books, girls</i>)	NN2	NN
following noun of title (e.g. <i>M.A.</i>)	NNA	NN
preceding noun of title (e.g. <i>Mr., Prof.</i>)	NNB	NN
singular locative noun (e.g. <i>Island, Street</i>)	NNL1	NN
plural locative noun (e.g. <i>Islands, Streets</i>)	NNL2	NN
numeral noun, neutral for number (e.g. <i>dozen, hundred</i>)	NNO	NN
numeral noun, plural (e.g. <i>hundreds, thousands</i>)	NNO2	NN
temporal noun, singular (e.g. <i>day, week, year</i>)	NNT1	NN
temporal noun, plural (e.g. <i>days, weeks, years</i>)	NNT2	NN
unit of measurement, neutral for number (e.g. <i>in, cc</i>)	NNU	NN
singular unit of measurement (e.g. <i>inch, centimetre</i>)	NNU1	NN
plural unit of measurement (e.g. <i>ins., feet</i>)	NNU2	NN
singular weekday noun (e.g. <i>Sunday</i>)	NPD1	NN
plural weekday noun (e.g. <i>Sundays</i>)	NPD2	NN
singular month noun (e.g. <i>October</i>)	NPM1	NN
plural month noun (e.g. <i>Octobers</i>)	NPM2	NN
base form of lexical verb (e.g. <i>give, work</i>)	VV0	VV
past tense of lexical verb (e.g. <i>gave, worked</i>)	VVD	VV
-ing participle of lexical verb (e.g. <i>giving, working</i>)	VVG	VV
-ing participle catenative (<i>going in be going to</i>)	VVGK	VV
infinitive (e.g. <i>work in It will work</i>)	VVI	VV
past participle of lexical verb (e.g. <i>given, worked</i>)	VVN	VV
past participle catenative (e.g. <i>bound in be bound to</i>)	VVNK	VV
-s form of lexical verb (e.g. <i>gives, works</i>)	VVZ	VV
general adjective (e.g. <i>old, good, strong</i>)	JJ	JJ
general comparative adjective (e.g. <i>older, better, stronger</i>)	JJR	JJ
general superlative adjective (e.g. <i>oldest, best, strongest</i>)	JJT	JJ
catenative adjective (<i>able in be able to</i>)	JK	JJ
adverb, after nominal head (e.g. <i>else, galore</i>)	RA	RR
adverb introducing appositional constructions (e.g. <i>namely</i>)	REX	RR
degree adverb (e.g. <i>very, so, too</i>)	RG	RR
wh- degree adverb (<i>how</i>)	RGQ	RR
wh-ever degree adverb (<i>however</i>)	RGQV	RR
comparative degree adverb (<i>more, less</i>)	RGR	RR
superlative degree adverb (<i>most, least</i>)	RGT	RR
locative adverb (e.g. <i>alongside, forward</i>)	RL	RR
prep. adverb, particle (e.g. <i>about, in</i>)	RP	RR
prep. adv., catenative (<i>about in be about to</i>)	RPK	RR
general adverb (e.g. <i>always, typically</i>)	RR	RR
wh- general adverb (<i>where, when, why, how</i>)	RRQ	RR
wh-ever general adverb (<i>wherever, whenever</i>)	RRQV	RR
comparative general adverb (e.g. <i>better, longer</i>)	RRR	RR
superlative general adverb (e.g. <i>best, longest</i>)	RRT	RR
quasi-nominal adverb of time (e.g. <i>now, tomorrow</i>)	RT	RR

Table 2: Part-of-speech tag reduction mapping.

SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek

Konstantin Schulz and Florian Deichsler

Humboldt University Berlin
konstantin.schulz@hu-berlin.de

Abstract

Literary scholars of Latin and Ancient Greek increasingly use natural language processing for their work, but many models and datasets are hard to use due to a lack of sustainable research data management. This paper introduces the Systematic Evaluation Framework for natural language processing models and datasets in Latin and Ancient Greek (SEFLAG), which consistently assesses language resources using common criteria, such as specific evaluation metrics, metadata and risk analysis. The framework, a work in progress in its initial phase, currently covers lemmatization and named entity recognition for both languages, with plans for adding dependency parsing and other tasks. For increased transparency and sustainability, a thorough documentation is included as well as an integration into the HuggingFace ecosystem. The combination of these efforts is designed to support researchers in their search for suitable models.

1 Introduction

Recent years have seen a surge of publications employing natural language processing (NLP) for the analysis of ancient texts (Papantoniou and Tzitzikas, 2020; Ehrmann et al., 2021; Sommerschild et al., 2023). However, as with other historical languages (see Zhou et al. (2023) for Classical Chinese), the communities around Latin and Ancient Greek rarely provide standardized and centralized resources specifically for the training and evaluation of NLP models. The corresponding treebanks in Universal Dependencies are one notable exception. Lemmatization, on the other hand, is notorious for the many different approaches to, e.g., character encoding (Tauber, 2019), handling of diacritics (Kostkan et al., 2023), homographs (Mambri and Passarotti, 2019), and other challenges. Besides, existing NLP models are scattered across many different technical platforms such as spaCy (Burns, 2023), Flair NLP (Yousef et al., 2023) or

Google Cloud (Bamman and Burns, 2020). As a consequence, every member of the Classics community has to collect and evaluate the same resources again. This leads us to the central question of how we can support literary scholars of Latin and Ancient Greek in choosing the right NLP models for their research agenda.

To address this challenge, we present SEFLAG¹, the Systematic Evaluation Framework for NLP models and datasets in Latin and Ancient Greek. Our work is still in progress, so we share only a small proof of concept with lemmatization and named entity recognition (NER). Next up, dependency parsing will follow. Our contributions are as follows:

- We collect and document existing datasets and NLP models, using recently established standards such as datasheets (Gebru et al., 2021) and model cards (Mitchell et al., 2019).
- We create benchmarks from suitable datasets, use consistent metrics for comparing models' performance on them and publish results in the Hugging Face² (HF) ecosystem.
- We document and publish conceptual mappings for connecting specific NLP models and datasets that were originally built using different annotation guidelines.

2 Related Work

Building highly specialized frameworks like SEFLAG can suffer from various problems. For example, transferring modern developments (large language models, analytical categories) to ancient contexts is non-trivial (McGillivray, 2013; Singh et al., 2021; Ehrmann et al., 2021; Riemenschneider and Frank, 2023; Yousef et al., 2023). In particular, the loose distinction between different seg-

¹<https://github.com/daidalos-project/seflag>

²<https://huggingface.co/>

ments of a text in Vedic Sanskrit (Biagetti et al., 2021) and other historical languages necessitates elaborate interpretative efforts to introduce modern syntactic concepts like punctuation or main and subordinate clause. This relatively relaxed notion of syntax correlates with the rather pronounced linguistic variation of Latin and Ancient Greek due to their diachronic, diatopic and diastratic differences (Kostkan et al., 2023).

Similar issues arise in the treatment of historical newspapers (Ehrmann et al., 2020) and early modern scientific texts (Odebrecht et al., 2017), which indicates a general trend of higher linguistic variation and lesser availability of language resources for historical languages (Etxeberria et al., 2016). For ancient texts, there is the additional burden of manifold textual transmission (including indirect transmission through citations), which prevents us from establishing texts in their original form with certainty (Sommerschild et al., 2023).

Besides, existing NLP implementations for textual annotation often do not fully adhere to the FAIR (Wilkinson et al., 2016) guiding principles of research data management. Earlier evaluations of available resources, on the other hand, were often performed indirectly, e.g., by carrying out surveys in the user community (Monachini et al., 2018) rather than directly testing the resources. Finally, large-scale data processing necessitates automation due to its efficiency, but automation can lead to a loss of data quality in highly heterogeneous datasets (Passarotti and Mambrini, 2022) and is often not sufficient for unifying multiple conceptually different resources, e.g., for valency patterns in ancient languages (Luraghi et al., 2024).

Fortunately, there has also been some progress in NLP for Latin and Ancient Greek. Most researchers involved in NLP evaluation choose one of the two languages and a single NLP task, like word embeddings (Stopponi et al., 2023) or topic modeling (Martinelli et al., 2024). A few of them even work on both languages, usually for single tasks that can be addressed through inherently cross-lingual methods (Perrone et al., 2021). Some engineering work has been done by the CLTK team (Johnson et al., 2021), offering a solid basis for data processing in historical languages but still suffering from a lack of evaluation and benchmarking. Nevertheless, all of these approaches are an important foundation for our mission of collecting and disseminating such resources in a centralized

manner.

A more coordinated and comprehensive initiative was the LiLa project³, which managed to successfully collect, harmonize and disseminate multiple existing language resources for Latin (Mambrini et al., 2020). In particular, they launched the EvaLatin evaluation campaign (Sprugnoli et al., 2020). Unfortunately, their project has officially ended and their platform (which is still running) does not cover Ancient Greek at all. Even for Latin, it addresses many NLP tasks, but not all: NER, topic modeling and some others are missing.

Other platforms have partly solved the problem of long-term availability and funding, such as the Perseus Digital Library. Like LiLa, they use Linked Data (Almas et al., 2014) to make their content findable and interoperable, but do not support evaluation reports. Generic language platforms like HF offer such reporting, but cater to a different audience (namely computer scientists and computational linguists), thus neglecting our target group of literary scholars.

3 Methodology

Our intermediate goal is to find existing NLP models and apply them to existing datasets (both having been created by others). Then, we perform one measurement for each of the two languages: Predictions of the NLP model are assessed using the ground truth annotations from the dataset and consistent metrics. For enhanced transparency and reusability, we document conceptual mappings, as most datasets and NLP models were created using rather different annotation schemas.⁴

For example, names of fictional characters in ancient literature may count as *PERSON* names in one NER dataset, but not in others (see section 4). This results in many datasets and NLP models that belong to the same task (namely, NER), but are not easily interoperable. Conceptual bridging (e.g., through mappings) is needed to close this gap and enable combinations of those resources. Even

³<https://lila-erc.eu>

⁴The problem is well-known especially in the treebanking community, where the de-facto standard of Universal Dependencies has been the most prominent effort to harmonize various other existing traditions such as the Index Thomisticus Treebank (Cecchini et al., 2018) or the Latin Dependency Treebank (Bamman and Crane, 2011). Such issues are particularly pressing for low-resource languages like Latin and Ancient Greek, where data sparsity hampers the development and application of various NLP technologies (McGillivray, 2013). In these languages, pushing the boundaries of existing resources by making them interoperable is especially important.

language resources that do not share the same annotation schema may still profit from the unification of certain annotation labels (such as *PERSON* and *PRS*, see section 4), depending on their conceptual overlap.

For enhanced sustainability, in applicable cases like lemmatization, we merge multiple datasets (i.e., various treebanks from Universal Dependencies) into a larger benchmark and publish it as a HF Dataset⁵ for the corresponding task. In doing so, we adopt the approach of Sprugnoli et al. (2020) by integrating diverse language material with regard to time and genre.

Additionally, we use datasheets and model cards (see Appendix A and B) to describe language resources systematically. Ideally, such datasheets and model cards should be provided by the creators themselves. However, the adoption of those standards is still insufficient in the Classics community. As the next best option, we create such materials ourselves and try to infer their content from publicly available information about the resources (in scientific publications, source code repositories, etc.). They will be uploaded to the HF Space⁶ of an NLP infrastructure (see section 4) and integrated into their website. This infrastructure allows our users to directly apply the evaluated NLP models to their own datasets and learn more about the various tasks through open educational resources.

Model cards are provided separately for each NLP model in our evaluation. They include general metadata like license, version or architecture, but also more complex considerations like ethical implications, ecological factors and possible risks of certain use cases. Apart from literary scholars of Latin and Ancient Greek, we also take neighboring disciplines into account, such as historians, theologians or archeologists dealing with ancient textual materials. From our point of view, their shared characteristics are limited technical background knowledge (Caraher, 2020) and a high interest in practical applications as well as methodological innovation, though all of these aspects are somewhat disputed in the scientific literature (Buchanan, 2015; Mahony, 2016; Damer, 2023).

⁵https://huggingface.co/datasets/daidalos-project/latin_treebanks_ud_test

⁶<https://huggingface.co/daidalos-project>

4 Implementation

For long-term sustainability, we aim to integrate our work into the Daidalos research infrastructure⁷ with institutionalized governance as provided by the datacenter⁸ of Humboldt University Berlin⁹, which offers a dedicated cloud computing service (as recommended by Almas (2017)). Funding for such an infrastructure has already been secured for an initial period of 3 years, which can be extended to about 10 years depending on periodic evaluation. As a consequence of our integration into that infrastructure, we also build on their community work: Their already existing biannual workshops, national research partnerships with classical scholars and open educational resources on Historical Language Processing¹⁰ are the backbone of our strategy to interact with our target audience and disseminate the evaluation results as widely as possible.

We provide explicit mappings for two cases¹¹: NLP models that are evaluated internally (on the test split of their original training data), and externally (i.e., a completely new dataset). The internal case is covered by the *flair_grc_multi_ner* tagger being tested on the data¹² that was curated by Yousef et al., i.e., a mixture of Herodotus, Homer and Athenaeus of Naucratis. Under those circumstances, no mapping is needed at all because the model was conceptualized directly with that dataset in mind.

Mapping the external case is more challenging: We took the LatinCy model and applied it to the Herodotos Project¹³ dataset. Each of the two resources uses 4 different entity tags that roughly correspond to the original ones introduced in Grishman and Sundheim (1996). The *PERSON* and *PRS* classes are arguably most compatible. However, the annotation guidelines for neither of the two

⁷<https://daidalos-projekt.de>

⁸<https://www.cms.hu-berlin.de/en/>

⁹<https://www.hu-berlin.de/en>

¹⁰<https://daidalos-projekt.de/dokumentation>

¹¹See <https://github.com/daidalos-project/seflag/blob/main/mappings.yaml>. To the best of our knowledge, there is no existing best practice for documenting linguistic annotation mappings. In particular, different conversion software like Pepper (Zipser and Romary, 2010) or Grew (Guillaume, 2021) uses different data formats for serializing the respective conversion instructions.

¹²https://github.com/daidalos-project/seflag/blob/main/documentation/datasheets/yousef_et_al_dataset.md

¹³<https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation>

language resources have been published anywhere. Thus, we cannot say for sure if their rules for assigning labels to named persons match each other, even considering the vague statement in Burns (2023) that the label applies to "people, including fictional". As a last resort, since the Herodotos dataset was included as training data for the Lat-inCy model, we may conclude that all 3 entity classes roughly correspond to each other, which allows us to apply mappings and pair the two for evaluation.

5 Results

We report the evaluation results in Table 1. We choose macro F1 and accuracy because they are applicable to many use cases and are widely adopted in the scientific community¹⁴. Moreover, macro F1 can be indicative of certain characteristics of language resources (Bone et al., 2015) such as the balance of the data distribution. We believe that it is part of our mission to inform potential users about weaknesses in a dataset, such as the strong class imbalance in both NER datasets (see Appendix B): Since most words in a text are non-entities, it is easy to achieve high accuracy by always guessing 'non-entity' as a baseline. This also explains the comparatively low scores in our NER evaluation, where non-entities are treated as rather unimportant. Furthermore, we will publicly upload our results to the HF Hub, so many others can benefit from the insights and do not have to run the evaluations themselves. This saves time and resources for the research community, while also providing easier access to necessary information about language resources.

A qualitative analysis of the lemmatization results empirically reveals some of the problems that were outlined in the research literature (see section 1): Variant spellings of the same lemma exist due to flexible orthography (οὔτως or οὔτω(ς), *parvulus* or *paruulus*), capitalization (*Romanus* or *romanus*), diacritics (τίς, τῖς or τῆς), and separate entries for specific inflected forms (χύκλως or χύκλωφ, *diuerto* or *diuersus*).

6 Limitations and Risks

Our approach of curating datasheets and model cards for resources that we did not create ourselves leads to information gaps in the documentation.

¹⁴For examples from the Classics, see Bizzoni et al. (2014); Stoeckel et al. (2020); Köntges (2020).

Nevertheless, our effort of inferring information from other sources and disseminating it in a centralized, systematic fashion is highly beneficial for the targeted research community.

Currently, we only report rather simple metrics. To enable deeper insights into model behavior and dataset structure, we plan to add class-wise confusion matrices, detailed qualitative error analyses and task-specific metrics (like the ones introduced by the Message Understanding Conference (Nadeau and Sekine, 2007)) which allow to distinguish between errors related to entity status and entity type.

Finally, we are very confident that our framework scales well to other planned tasks like part-of-speech tagging, sentiment analysis and dependency parsing.¹⁵ As an abstract representation of the different linguistic annotations, for example, we intend to use a graph model like SALT (Zipser and Romary, 2010). Besides, we have clear interfaces for adding more models, datasets and evaluation metrics. As of now, however, it is unclear to which extent we may need to introduce further metadata (spatial, temporal, stylistic, etc.).

7 Conclusions

Our evaluation framework SEFLAG aims to support literary scholars of Latin and Ancient Greek in selecting the right NLP models for their research. We provide quantitative evaluations of existing models on suitable datasets. Conceptual mappings between tagsets used for the annotation of different language resources are documented explicitly and in a human-readable way. Evaluation results are reported using common metrics (F1, accuracy) and are accompanied by additional documentation for the language resources: datasheets for datasets and model cards for NLP models. Using that additional information, we enable researchers to critically assess the value of such resources for their own research, including questions of dataset characteristics, model architecture, annotation guidelines and contact persons. In short, we provide low-level access to the costly and complex task of NLP evaluation for Latin and Ancient Greek, with a proof of concept that focuses on NER and lemmatization.

In the near future, we will work on fully integrating our framework into the Daidalos research

¹⁵These tasks have been chosen because they are of general interest to the research community (Berti, 2019; Ehrmann et al., 2021; Beersmans et al., 2023) and are directly relevant to the Daidalos research infrastructure.

Language	Task	Model	Dataset	Metric	Score
Latin	NER	LatinCy	Herodotos	macro F1 ↑	58
	lemmatization	LatinCy	UD Latin	accuracy ↑	88
Ancient Greek	NER	flair_grc_bert_ner	Yousef et al.	macro F1 ↑	64
	lemmatization	greCy	UD Ancient Greek	accuracy ↑	89

Table 1: Evaluation results for NER and lemmatization in Latin and Ancient Greek. The metrics used macro F1 and accuracy. 3 different NLP models have been evaluated on 4 different datasets. Upward arrows for a metric indicate that higher scores are better.

infrastructure. Furthermore, we would like to add more NLP tasks, models and datasets. Finally, we also want to create Open Educational Resources to educate interested researchers about central essentials of the evaluation, such as specific metrics, task concepts and annotation approaches.

Ethics Statement

We address ethical considerations mainly through heavy use of model cards and datasheets. Besides, we respect licensing conditions for datasets by publishing our benchmarks only if all contained sub-datasets allow it from a legal perspective, and only under a license that matches the ones used in the sub-datasets. In general, we refrain from reusing datasets with licenses that are too prohibitive.

Acknowledgements

We are grateful to five anonymous reviewers and Andrea Beyer as well as Anke Lüdeling for commenting on earlier drafts of this paper.

This work is part of a project funded by the German Research Foundation (project number 518919950) and led by Andrea Beyer, Malte Dreyer and Anke Lüdeling.

References

- Bridget Almas. 2017. Perseids: Experimenting with Infrastructure for Creating and Sharing Research Data in the Digital Humanities. *Data Science Journal*, 16(19):1–17.
- Bridget Almas, Alison Babeu, and Anna Krohn. 2014. Linked Data in the Perseus Digital Library. *ISAW Papers*, 7(3).
- David Bamman and Patrick J Burns. 2020. [Latin BERT: A Contextual Language Model for Classical Philology](#). *arXiv preprint arXiv:2009.10053*, pages 1–14.
- David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin Dependency Treebank. In *Language Technology for Cultural Heritage*, pages 79–98. Springer.
- Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training and Evaluation of Named Entity Recognition Models for Classical Latin. In *Ancient Language Processing Workshop*, pages 1–12.
- Monica Berti. 2019. Named entity annotation for ancient greek with inception. In *Proceedings of CLARIN Annual Conference*, pages 1–4.
- Erica Biagetti, Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2021. Evaluating Syntactic Annotation of Ancient Languages: Lessons from the Vedic Treebank. *Old World: Journal of Ancient Africa and Eurasia*, 1(1):1–32.
- Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory R Crane. 2014. The Making of Ancient Greek WordNet. In *LREC*, volume 2014, pages 1140–1147.
- Daniel Bone, Matthew S Goodwin, Matthew P Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. 2015. Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of autism and developmental disorders*, 45(5):1121–1136.
- Sarah A Buchanan. 2015. The Emerging Tradition of Digital Classics. *Annual Review of Cultural Heritage Informatics: 2014*, page 149.
- Patrick J Burns. 2023. [LatinCy: Synthetic Trained Pipelines for Latin NLP](#). *arXiv preprint arXiv:2305.04365*.
- William Caraher. 2020. Dissecting Digital Divides in Teaching. In Sebastian Heath, editor, *DATAM: Digital Approaches to Teaching the Ancient Mediterranean*, number 16 in Digital Press Books, pages 71–82. The Digital Press at the University of North Dakota.
- Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36.
- Erika Zimmermann Damer. 2023. What Is a Future for Classics? *American Book Review*, 44(3):47–50.

- Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*.
- Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: The Impresso Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968.
- Izaskun Etxeberria, Inaki Alegria, Larraitz Uria, and Mans Hulden. 2016. Evaluating the noisy channel model for the normalization of historical texts: Basque, Spanish and Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1064–1069.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for Datasets](#). *Communications of the ACM*, 64(12):86–92.
- Ralph Grishman and Beth M Sundheim. 1996. Design of the MUC-6 evaluation. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop Held at Vienna, Virginia, May 6-8, 1996*, pages 413–422.
- Bruno Guillaume. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.
- Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. [The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.
- Thomas Köntges. 2020. [Measuring philosophy in the first thousand years of Greek literature](#). *Digital Classics Online*, pages 1–23.
- Jan Kostkan, Márton Kardos, Jacob Palle Bliddal Mortensen, and Kristoffer Laigaard Nielbo. 2023. OdyCy—A general-purpose NLP pipeline for Ancient Greek. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–134.
- Silvia Luraghi, Alessio Palmero Arosio, Chiara Zanchi, and Martina Giuliani. 2024. Introducing PaVeDa—Pavia Verbs Database: Valency Patterns and Pattern Comparison in Ancient Indo-European Languages. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 79–88.
- Simon Mahony. 2016. [Open Education and Open Educational Resources for the Teaching of Classics in the UK](#). In Matteo Romanello and Gabriel Bodard, editors, *Digital Classics Outside the Echo-Chamber*, pages 33–50. Ubiquity Press.
- Francesco Mambrini, Flavio Massimo Cecchini, Greta Franzini, Eleonora Litta, Marco Carlo Passarotti, and Paolo Ruffolo. 2020. LiLa: Linking Latin. *Risorse linguistiche per il latino nel Semantic Web. Umanistica Digitale*, 4(8):63–78.
- Francesco Mambrini and Marco Passarotti. 2019. Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 71–80.
- Ginevra Martinelli, Paola Impiccihé, Elisabetta Fersini, Francesco Mambrini, and Marco Passarotti. 2024. Exploring Neural Topic Modeling on a Classical Latin Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6929–6934.
- Barbara McGillivray. 2013. *Methods in Latin Computational Linguistics*. Brill, Leiden; Boston.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.
- Monica Monachini, Anika Nicolosi, and Alberto Stefanini. 2018. Digital Classics: A Survey on the Needs of Ancient Greek Scholars in Italy. In *Proceedings of the CLARIN 2017 Conference*, pages 1–5. Linköping University Electronic Press.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.
- Carolin Odebrecht, Malte Belz, Amir Zeldes, Anke Lüdeling, and Thomas Krause. 2017. [RIDGES Herbiology: Designing a diachronic multi-layer corpus](#). *Language Resources and Evaluation*, 51(3):695–725.
- Katerina Papantoniou and Yannis Tzitzikas. 2020. NLP for the Greek language: A brief survey. In *11th Hellenic Conference on Artificial Intelligence*, pages 101–109.
- Marco Passarotti and Francesco Mambrini. 2022. Issues in Building the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *LLOD Approaches for Language Data Research and Management LLODREAM2022 : International Scientific Interdisciplinary Conference*.
- Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2021. [Lexical semantic change for](#)

- Ancient Greek and Latin.** In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*, number 6 in Language Variation, pages 287–310. Language Science Press.
- Frederick Riemenschneider and Anette Frank. 2023. **Exploring Large Language Models for Classical Philology.** In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 15181–15199.
- Pranaydeep Singh, Gorik Ruppen, and Els Lefever. 2021. **A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek.** In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.
- Thea Sommerschild, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.
- Rachele Sprugnoli, Marco Passarotti, Flavio Mas-similiano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvalLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).
- Manuel Stoeckel, Alexander Henlein, Wahed Hemati, and Alexander Mehler. 2020. Voting for POS tagging of Latin texts: Using the flair of FLAIR to better ensemble classifiers by example of Latin. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 130–135.
- Silvia Stopponi, Nilo Pedrazzini, Saskia Peels-Matthey, Barbara McGillivray, and Malvina Nissim. 2023. Evaluation of Distributional Semantic Models of Ancient Greek: Ancient Language Processing. *Proceedings of the Ancient Language Processing Workshop*, pages 49–58.
- James K Tauber. 2019. Character encoding of classical languages. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 10:137–158.
- Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. **The FAIR Guiding Principles for scientific data management and stewardship.** *Scientific Data*, 3(160018):1–9.
- Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2023. **Transformer-Based Named Entity Recognition for Ancient Greek.** In *Book of Abstracts*, pages 420–422, Graz. Zenodo.
- Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP Evaluation Benchmark For Classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319.
- Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*.

A Model Card: Latincy

la_core_web_lg

- Person or organization developing model: Patrick J. Burns; with Nora Bernhardt [ner], Tim Geelhaar [tagger, morphologizer, parser, ner], Vincent Koch [ner]
- Model date: May 2023
- Model version: 3.7.4
- Model type: spaCy
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features: For information on the training workflow see p.4-5 of LatinCy: Synthetic Trained Pipelines for Latin NLP (<https://arxiv.org/pdf/2305.04365v1>)
- Paper or other resource for more information: **Burns, P.J. 2023. "LatinCy: Synthetic Trained Pipelines for Latin NLP." arXiv:2305.04365 [cs.CL]. <http://arxiv.org/abs/2305.04365>.
- License: MIT

- Where to send questions or comments about the model: <https://diyclassics.github.io/>
- Intended Use
 - Primary intended uses: Morphological analysis, POS-Tagging, Lemmatizing, Parsing, NER
 - Primary intended users: Classical Scholars
 - Out-of-scope use cases: unknown
- Data, Limitations, and Recommendations
 - Data selection for training: Training data consists of latin UD-Treebanks, Wikipedia and OSCAR sentence data, the CC-100 Latin dataset and the Herodotos Project NER dataset
 - Data selection for evaluation: Evaluation was done according to the spaCy workflow and is documented in the meta.json file found in the repository (https://huggingface.co/latincy/la_core_web_lg/blob/main/meta.json)
 - Limitations: unknown

B Datasheet: Herodotos Project Dataset

For what purpose was the dataset created?

Was there a specific task in mind? Was there a specific gap that needed to be filled?

Please provide a description. created for Herodotos Project to train NER-Tagger (BiLSTM-CRF; see: Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeaux-Prunel and Marie-Catherine de Marnette. 2019. "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities." In Proceedings of North American Association of Computational Linguistics (NAACL 2019). Minneapolis, Minnesota.); Goal of Herodotos Project: catalogue and compendium of ancient ethnic groups; For more info on the corpus see: <https://aclanthology.org/W16-4012.pdf>

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

from the documentation: „The data files in the Annotation directory were annotated for named entities by a team of Classics experts at Ohio State University. Texts presently included are excerpts from Caesar’s Wars, both Gallic (GW) and Civil (CW), the Plinies’ writings, both Elder and Younger, and Ovid’s Ars Amatoria. ”

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. unknown

Any other comments? No

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. Latin texts "Texts presently included are excerpts from Caesar’s Wars, both Gallic (GW) and Civil (CW), the Plinies’ writings, both Elder and Younger, and Ovid’s Ars Amatoria."

How many instances are there in total (of each type, if appropriate)? 146,066 words

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). sample of Latin literature (see previous answers), representative of Classical Latin literature, might not be representative of the entire Latin literature (time, geography)

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. Each instance consists of raw text data

Is there a label or target associated with each instance? If so, please provide a

description. NER Labels: PRS-B, PRS-I, GEO-B, GEO-I, GRP-B, GRP-I or 0; labels follow the BIO scheme; see also: <https://aclanthology.org/W16-4012.pdf>

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. Relationships are made explicit according to the BIO scheme

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. Text from Gallic War is split into test and train sets

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Naturally occurring repetitions of names in the texts

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate. The dataset is self-contained and can be downloaded from GitHub (<https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation/blob/master/README.md>)

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description. No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. If the dataset does not relate to people, you may skip the remaining questions in this section. The dataset contains descriptions of war.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. A number of ethnic groups from antiquity are referred to.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how. Only historical individuals

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description. Only historical individuals

Any other comments? No

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? The data consists of publicly available texts

If the data was reported by subjects or indirectly inferred/derived from other data,

was the data validated/verified? If so, please describe how. unknown

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? from the documentation: „All texts are in Latin taken from the Latin Library Collection (collected by CLTK) or the Perseus Latin Collection. "

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? unknown

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? <https://aclanthology.org/W16-4012.pdf> S. 87: "an undergraduate, a graduate, and a professor of Classics, each with at least 4 years of experience studying Latin"

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. If the dataset does not relate to people, you may skip the remaining questions in this section. unknown

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. not applicable

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. not applicable

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). not applicable

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. not applicable

Any other comments? No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section. The data was manually annotated for NEs.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. The data can be downloaded from: https://github.com/clmarr/Herodotos-beta/tree/f22fdd92b3318cfb8fc93b004b0947aea14ce9c2/Annotation_1-1-19

Any other comments? No

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

It has been used to train an NER-Tagger for Latin. See: <https://aclanthology.org/W16-4012.pdf> and https://github.com/alexerdmann/HER/blob/master/HER_NAAACL2019_preprint.pdf

https://github.com/clmarr/Herodotos-beta/tree/f22fdd92b3318cfb8fc93b004b0947aea14ce9c2/Annotation_1-1-19 <https://github.com/Herodotos-Project/Herodotos-Project-Latin-NER-Tagger-Annotation>

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. What (other) tasks could the dataset be used for? See: https://github.com/alexerdmann/HER/blob/master/HER_NAAACL2019_preprint.pdf

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. AGPL-3.0 license

Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms? Strong class imbalance (most tokens are non-entities)

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. unknown

Are there tasks for which the dataset should not be used? If so, please provide a description. No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. unknown

Any other comments? No

Any other comments? No

Distribution

Maintenance

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The data can be downloaded from: https://github.com/clmarr/Herodotos-beta/tree/f22fdd92b3318cfb8fc93b004b0947aea14ce9c2/Annotation_1-1-19

Who will be supporting/hosting/maintaining the dataset? from the documentation: "Contact ae1541@nyu.edu or any of the co-authors with questions regarding this repository."

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? ae1541@nyu.edu

Does the dataset have a digital object identifier (DOI)? No

Is there an erratum? If so, please provide a link or other access point. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)? new instances for the Ancient Greek language will be added in the future

When will the dataset be distributed? The data can be downloaded from: https://github.com/clmarr/Herodotos-beta/tree/f22fdd92b3318cfb8fc93b004b0947aea14ce9c2/Annotation_1-1-19

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. not applicable

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers. unknown

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description. unknown

Any other comments? No

A Two-Model Approach for Humour Style Recognition

Mary Ogbuka Kenneth¹, Foad Khosmood², Abbas Edalat¹

¹Algorithmic Human Development group, Department of Computing, Imperial College London, UK

²Computer Engineering Department, California Polytechnic State University, USA

m.kenneth22@imperial.ac.uk, foad@calpoly.edu, a.edalat@imperial.ac.uk

Abstract

Humour, a fundamental aspect of human communication, manifests itself in various styles that significantly impact social interactions and mental health. Recognising different humour styles poses challenges due to the lack of established datasets and machine learning (ML) models. To address this gap, we present a new text dataset for humour style recognition, comprising 1463 instances across four styles (self-enhancing, self-deprecating, affiliative, and aggressive) and non-humorous text, with lengths ranging from 4 to 229 words. Our research employs various computational methods, including classic machine learning classifiers, text embedding models, and DistilBERT, to establish baseline performance. Additionally, we propose a two-model approach to enhance humour style recognition, particularly in distinguishing between affiliative and aggressive styles. Our method demonstrates an 11.61% improvement in f1-score for affiliative humour classification, with consistent improvements in the 14 models tested. Our findings contribute to the computational analysis of humour in text, offering new tools for studying humour in literature, social media, and other textual sources.

1 Introduction

Humour recognition is a multidimensional task influenced by various theories and manifested through diverse styles. There are various humour theories, such as relief, incongruity, and superiority theories (Morreall, 2011, 2012; Scheel and Gockel, 2017). The relief theory highlights the role of humour in relaxation, while the incongruity theory suggests that we find something funny when we notice a mismatch or contradiction between what we expect in a situation and what actually happens. The superiority theory suggests that people may laugh at other people's misfortunes in an effort to demonstrate their superiority.

These theories not only explain why we find

things humorous but also why we laugh as a response. In recent decades, evolutionary psychology has introduced a new perspective on laughter itself, known as the *play* theory (Martin and Ford, 2018): laughter developed as a play signal in higher primates in their mock fights to indicate non-aggressive intent.

Laughter, therefore, is more than just a reaction to humour; it serves various functions, including promoting mental, emotional, and physical well-being. This idea forms the basis for laughter therapy, a cognitive-behavioural treatment designed to induce laughter and reduce stress, tension, anxiety, and sadness (Yim, 2016). However, as Martin et al. (2003) noted, not all humour is beneficial—some forms can even harm relationships with others or oneself.

Considering its impact on well-being, Martin et al. (2003) categorised humour into four styles: self-enhancing, self-deprecating, affiliative, and aggressive. Affiliative and self-enhancing humour are beneficial to psychological well-being. Affiliative humour fosters social bonding, while self-enhancing humour involves maintaining a positive outlook without harming oneself or others, often employed as a coping mechanism in difficult situations (Edalat, 2023; Kenneth et al., 2024; Hampes, 2007; Plessen et al., 2020). In contrast, aggressive and self-deprecating humour can be harmful. Aggressive humour, rooted in superiority theory, belittles or mocks others, whereas self-deprecating humour seeks approval by making oneself the target of jokes (Khrantsova and Chuykova, 2016; Kuiper et al., 2016; Veselka et al., 2010).

In artificial intelligence (AI), humour is considered AI-complete (Shani et al., 2021; Strapparava et al., 2011; Kenneth et al., 2024), meaning that a system capable of producing and recognising human-like humour would possess general intelligence. Despite the importance of humour, most computational efforts have focused on laughter de-

tection (Vargas-Quiros et al., 2023; Matsuda and Arimoto, 2023; Inoue et al., 2022), classification (Tanaka and Campbell, 2014) and generation (Inoue et al., 2022), as well as humour detection (Oliveira et al., 2020; Jaiswal et al., 2019; Chauhan et al.), and humour generation (Luo et al., 2019; He et al., 2019; Yu et al., 2018), with little emphasis on humour styles and their links to well-being. Kenneth et al. (2024) identified a gap in the current ML landscape: the lack of datasets and models specifically designed to recognise these four humour styles.

Building on the gaps identified by Kenneth et al. (2024), this study addresses the lack of an established dataset and ML models for recognising the four humour styles: self-enhancing, self-deprecating, affiliative, and aggressive. We draw on Martin et al. (2003), who defined and validated these styles, providing the theoretical basis for our classification task. Additionally, Edalat (2023)’s work on self-initiated humour protocols (SIHP) informs how different humour styles can enhance well-being, while Amjad and Dasti (2022) research on the link between humour styles, emotion regulation, and subjective well-being highlights the potential applications of our work in psychological and clinical contexts. By integrating these insights, we aim to develop a comprehensive approach to humour style recognition grounded in psychological theory and applicable to real-world scenarios. The key contributions of this paper are:

1. Introduction of a new text dataset for humour style recognition, addressing the lack of established datasets. This dataset is publicly available to the community.
2. Baseline evaluations using various ML classifiers and models.
3. Development of a two-model approach for improved humour style recognition.
4. Extensive evaluation of the proposed two-model approach.

2 Related Works

Humour recognition and classification are active research areas in NLP and multi-modal analysis. While our focus is on humour style recognition, we draw insights from related fields like general humour detection and sarcasm detection.

Weller and Seppi (2020) compiled a dataset of 550,000 jokes from Reddit posts, using user ratings and engagement metrics as quantifiable humour

quality measurements. However, the dataset’s reliance on Reddit data alone may introduce biases and limit generalisability. Our study addresses this by introducing a more diverse dataset specifically tailored for humour style recognition from various online sources.

Oliveira et al. (2020) explored humour recognition in Portuguese text, achieving a 75% f1-score using Naive Bayes, Support Vector Machine, and Random Forest classifiers. However, their work was limited to binary classification of headlines and one-liners. Our approach extends this by focusing on multi-class classification of humour styles in both short and long texts.

Tang et al. (2022) created a dataset and classification model for sub-types of inappropriate humour, using large language models like BERT. While relevant, their focus on inappropriate humour differs from our goal of recognising humour styles linked to psychological well-being.

Kamal and Abulaish (2020) targeted self-deprecating humour, one of the four styles we examine. Their use of specific feature categories (self-deprecating pattern, and word embedding) informs our feature engineering process. However, our study broadens the scope to include all four humour styles.

Christ et al. (2022a,b) developed models for humour recognition in German football press conferences. Although their work yielded promising results, it was limited to the MuSe humour challenge and the Passau-SFCH German dataset, unlike our broader approach.

Sarcasm detection is closely related to humour style recognition since it is often used in aggressive and self-deprecating humour styles. Liang et al. (2021) used an interactive graph convolution network for multi-modal sarcasm detection, highlighting the importance of contextual cues. This technique could be adapted to distinguish humour styles.

Jinks (2023) improved sarcasm detection with a two-step fine-tuning process using RoBERTa, a method that could enhance humour style classification given the subtle differences between styles.

Fang et al. (2024) introduces the Single-Stage Extensive Semantic Fusion model for multi-modal sarcasm detection by concurrently processing and fusing multi-modal inputs in a unified framework. This approach could be adapted for humour style recognition, when we expand our dataset to include multi-modal features in the future.

Although these studies contribute to the detection of humour and sarcasm, there is a gap in recognising the four humour styles defined by Martin et al. (2003). Our work fills this gap by creating a dedicated dataset and developing classification models tailored to these humour styles.

3 Dataset Collection and Annotation

A significant challenge in identifying humour styles automatically is the lack of annotated datasets suitable for training machine learning models. To address this, we created a comprehensive dataset comprising 1,463 instances from various sources:

1. 983 jokes from several well-known websites where jokes were labelled by users or editors.
2. 280 non-humorous text instances from the ColBERT dataset (Annamoradnejad and Zoghi, 2020).
3. 200 instances from the Short Text Corpus¹, consisting of 150 jokes and 50 non-jokes

After annotation, the dataset consists of 298 instances of self-enhancing humour, 265 of self-deprecating humour, 250 of affiliative humour, 318 of aggressive humour, and 332 neutral instances, with text lengths ranging from 4 to 229 words. This distribution ensures balanced representation across the different humour styles and neutral text.

3.1 Data Sources and Labelling

The 983 jokes were extracted from sources like Reader’s Digest, Parade, Bored Panda, Laugh Factory, Pun Me, Independent, Cracked, Reddit, Tastelessly Offensive and BuzzFeed. We labelled each joke based on the original labels, definitions, or tags given on the websites, mapping them to our categories based on humour theory. Table 1 summarises these mappings, illustrating how the website tags correspond to our humour style labels.

Equivalence Classes (Website Keywords)	Humour Styles
Dark (inappropriate) Jokes	Aggressive
Insult	Aggressive
Icebreakers Jokes for Work Meetings	Affiliative
International Day of Happiness	Affiliative
Friendship	Affiliative
Family jokes	Affiliative
Classroom jokes	Affiliative
Self-deprecating	Self-deprecating
Self-love	Self-enhancing
Self-care	Self-enhancing

Table 1: Terminological Equivalence Classes

¹Short Text Corpus (<https://github.com/CrowdTruth/Short-Text-Corpus-For-Humor-Detection>)

Humour Styles Examples	
Self-enhancing joke examples: 1. The bags under my eyes are Prada. 2. Both laughing and crying are reactions to irritation and tiredness. I personally like to laugh more because there is less cleanup required afterwards. 3. I am not pretty but also pretty awesome, pretty smart, and pretty kind. 4. My sense of humour makes the world a better place.	Affiliative joke examples: 1. You look like what John Legend sounds like. 2. I don't know why, but I find it weirdly alluring when you do embarrassing things. 3. Accept who you are. Unless you are a serial Killer. 4. You are even better than unicorn because you are real.
Self-deprecating joke examples: 1. I may be trash, but I burn with a bright flame 2. I don't have a nervous system. I am a nervous system! 3. I am clumsy that I once tripped over my own shadow. 4. My fake plants died because I forgot to pretend to water them.	Aggressive joke examples: 1. I hear you were born on a farm. Any more in the litter? 2. Shock me say something intelligent. 3. I can't talk to you right now, tell me, where will you be in 10 years? 4. Behind every fat woman there us a beautiful woman. No seriously, your in my way.

Figure 1: Joke Examples for Each Humour Style

For example, in Table 1 the "Dark (inappropriate)" tag was mapped to the aggressive style because dark or inappropriate jokes are identified as being cruel, morbid, or offensive to some, which aligns with the characteristics of aggressive humour (Tang et al., 2022). Further details on these mappings are available in Appendix B.

To simulate real-life scenarios where users might input non-humorous text, we added 280 non-humorous instances from the ColBERT dataset (Annamoradnejad and Zoghi, 2020), labelled as Neutral.

Figure 1 presents random examples from the dataset for each humour style. Additionally, word clouds showing the most common words associated with each humour style in the created dataset are provided in Appendix C.

3.2 Dataset Composition and Potential Biases

Each humour style in our dataset was primarily sourced from different websites (see Table 11 in Appendix A for details). The use of diverse websites, catering to various audiences and content styles, helps mitigate biases that could arise from relying on a single source. However, since the jokes were collected in English, there may be language biases, as humour often involves nuances and idioms specific to certain languages and cultures.

By aggregating data from multiple websites, we aimed to reduce inherent biases from any single source and provide comprehensive coverage of different humour styles, enhancing the robustness of the dataset. However, most websites (except Reader’s Digest and Laugh Factory) featured jokes from only one humour type, potentially introducing idiosyncratic styles that could lead the classifier to learn spurious correlations.

To address this concern and further diversify our dataset, we included an additional 200 jokes from

the existing Short Text Corpus joke dataset¹ and have them annotated by six human annotators. Details of the Short Text Corpus¹ and the annotation process are discussed further in the following subsection.

3.3 Annotation Process and Inter-annotator Agreement

Building on our efforts to address potential biases in our dataset composition, we took additional steps to ensure the robustness of our data. To mitigate potential biases from idiosyncratic styles of the individual websites, we randomly selected 200 instances from the Short Text Corpus¹, dividing them into two sets of 100 samples. This corpus was chosen for its diversity, featuring both short and long jokes from more than seven sources, as well as non-jokes from three sources. In contrast, the ColBERT dataset (Annamoradnejad and Zoghi, 2020) was not used here because it consists solely of Reddit jokes, which would not address the issue of spurious correlations.

To further ensure the reliability of our annotations, we recruited six Ph.D. candidates from Africa, Asia, and Europe to serve as annotators, bringing a diverse range of analytical perspectives to the task. Each set of 100 samples was independently annotated by three annotators, who were provided with humour style definitions and asked to classify each instance as self-enhancing, self-deprecating, aggressive, affiliative, or neutral. A majority vote determined the final label for each instance.

Fleiss’ Kappa was used to assess inter-annotator agreement. The results showed fair agreement levels:

1. First 100 samples: Fleiss’ Kappa = 0.2651
2. Second 100 samples: Fleiss’ Kappa = 0.2841

Despite the relatively low Kappa values, further analysis showed substantial agreement among at least two annotators:

1. For the first set of 100 samples: 91 samples had at least two annotators agreeing on the label and 9 instances had all three annotators disagreeing.
2. For the second set of 100 samples: 95 samples had at least two annotators agreeing on the label and 5 instances had all three annotators disagreeing.

To resolve the 14 instances (9 in the first set, 5 in the second) where all three annotators disagreed, indicating no majority vote, we used four

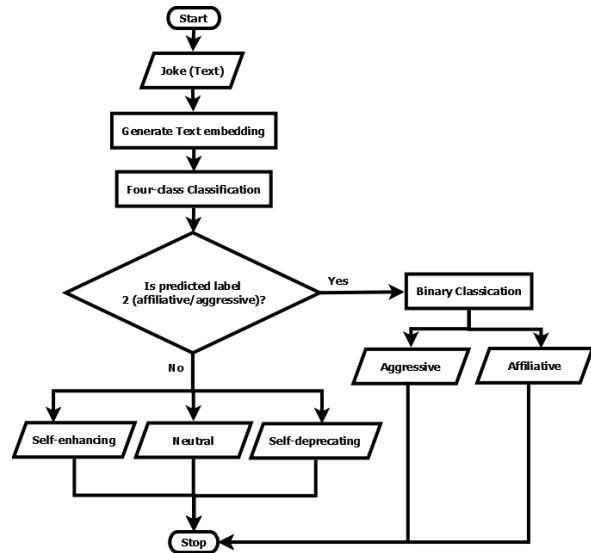


Figure 2: Flowchart illustrating the proposed Two-Model Approach for Humour Style Recognition

Large Language Models (LLMs) chatbots: [ChatGPT](#), [Claude](#), [Microsoft Copilot](#), and [HuggingChat](#) - to classify the jokes. We prompted the LLMs to categorise each joke instance as self-enhancing, self-deprecating, aggressive, affiliative, or neutral. Each of the 14 instances then had seven labels (from the 4 LLMs and 3 human annotators), and the majority label was assigned. Table 12 in Appendix D provides examples of instances where annotators disagreed, along with the annotators’ and LLMs’ labels.

These disagreements highlight the subjective nature of humour interpretation, which can be influenced by cultural differences, personal experiences, and individual preferences (Lu, 2023). This subjectivity is a natural aspect of humour annotation, and our use of multiple annotators and LLMs helps to mitigate its impact.

4 Methodology

This study employs two different approaches for humour style recognition: the single-model and the two-model approach. A total of 14 models were evaluated, including Naive Bayes, Random Forest, XGBoost (each with six different text embeddings), and DistilBERT. Figure 2 illustrates the two-model approach, which first classifies humour instances into broader groups before refining to specific styles.

4.1 Classifiers and Embedding Models

4.1.1 Classifiers

The selection of classifiers was based on their suitability for the task at hand and efficiency in low-resource settings, avoiding resource-intensive large language models such as GPT4 and LLaMA prone to overfitting on small datasets due to their complex architectures (Schur and Groenjes, 2024; Diwakar and Raj, 2024; Berfu B et al., 2020):

Naive Bayes (NB): A probabilistic classifier based on the Bayes Theorem, assuming conditional independence of features given the target class (Berrar, 2019).

Random Forest (RF): A bagging ensemble classifier using majority voting from multiple decision trees (Jin, 2020).

eXtreme Gradient Boosting (XGBoost): A boosting ensemble classifier aggregating predictions of several weak learners, with regularisation to prevent overfitting (Jiang et al., 2019).

DistilBERT: A condensed BERT variant, offering faster performance and memory efficiency while maintaining competitive performance on NLP tasks (Sanh et al., 2019).

4.1.2 Sentence Embedding Models

To capture distinct linguistic nuances and improve classification performance, we selected six embedding models from the top 20 on the Massive Text Embedding Benchmark (MTEB) leaderboard. These models were chosen for their robustness, efficiency, speed, and lightweight memory usage:

- General Text Embeddings (GTE) and GTE Upgraded (ALI) (Li et al., 2023)
- BAAI General Embedding (BGE) (Xiao et al., 2022; Zhang et al., 2023)
- Matryoshka Representation Learning and Binary Quantization (MRL) (Lee et al., 2024)
- Universal AnglE Embedding (UAE) (Li and Li, 2023)
- Multilingual E5 Text Embeddings (MUL) (Wang et al., 2024)

These embeddings were combined with RF and XGBoost classifiers for humour style recognition.

4.2 Single-Model Approach

In this approach, a single ML model is trained to classify the input text into one of the five classes: self-enhancing (**label 0**), self-deprecating (**label 1**), affiliative (**label 2**), aggressive (**label 3**), and neutral (**label 4**). This approach treats the task

as a multi-class classification problem, where the model needs to distinguish between all five classes simultaneously.

To provide insight into the single-model performance, Figure 3 presents the confusion matrices for the 5-fold cross-validation results of four models: Naive Bayes (NB) 3a, GTE+RF 3b, MUL+XGBoost 3c, and UAE+RF 3d.

4.3 Two-Model Approach

To address limitations observed in the single-model approach, particularly in distinguishing affiliative humour, we developed a two-model approach. This method, inspired by previous studies (Khan et al., 2022; Van Lam et al., 2011; Demidova, 2021), improves classification performance by breaking down the problem into multiple steps.

The rationale behind this approach is to first separate the instances into broader groups and then focus on the more challenging task of distinguishing between affiliative and aggressive humour styles. This strategy is informed by an analysis of misclassified samples from the cross-validation and test set evaluation of the single-model approach, which revealed that affiliative humour was predominantly misclassified as aggressive humour. This pattern of misclassification is clearly illustrated in the cross-validation confusion matrices shown in Figure 3.

The two-model approach involves two sequential steps:

1. **Step 1: Four-Class Classification Model:** Train an ML model to distinguish between self-enhancing, self-deprecating, neutral, and a combined affiliative/aggressive class.
2. **Step 2: Binary Classification Model:** Train a separate binary classification model to distinguish between affiliative and aggressive instances from the combined class in step 1.

This approach allows for optimising overall performance by combining the best-performing models for each subtask.

4.4 Experimental Setup

The humour styles dataset was split 80/20 for training and testing, randomised using a fixed seed of 100 to ensure reproducibility. We used 5-fold cross-validation for all experiments to validate model performance and prevent overfitting. For the NB classifier, we used a smoothing parameter of 1. The RF and XGBoost classifiers were implemented using their default hyperparameters. The DistilBERT

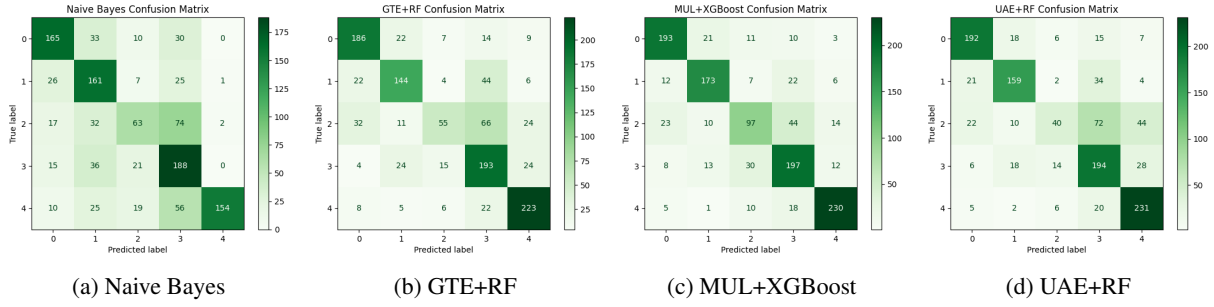


Figure 3: 5-Fold Cross Validation Confusion Matrix

model was fine-tuned for 5 epochs with a weight decay of 0.01, warmup steps of 500, and a training batch size of 8, using the default learning rate scheduler provided by the Hugging Face Transformers library.

4.5 Evaluation Metrics

Model performances were evaluated using standard metrics: accuracy, precision, recall, and f1-score. Accuracy measures overall performance, precision quantifies the ratio of true positives to predicted positives, recall assesses the model’s ability to identify actual positives, and f1-score represents the harmonic mean of precision and recall. Furthermore, the Wilcoxon signed-rank test was used to compare the single-model and two-model approaches, determining the statistical significance of the performance differences between these approaches.

5 Results and Discussions

Experiments for the single-model and two-model approaches were conducted on Fourteen models: NB, RF + six embedding models, XGBoost + six embedding models and DistilBERT.

5.1 Baseline Model (Single-Model Approach)

Tables 2 and 3 show the mean accuracy and macro-mean f1-score of the 5-fold cross-validation for different models and embedding techniques, respectively. The results highlight the robustness and generalisability of our models across different data splits.

Table 4 presents the overall performance for the five-class classification. MUL+RF, ALI+RF, and DistilBERT performed best with accuracies and f1-scores of 77.1% and 76.6%, 77.8% and 77.3%, and 75.4% and 75.2%, respectively.

While the single-model approach achieved decent overall performance, Table 5 reveals that all

models struggle to identify affiliative humour accurately. Despite high overall accuracy, this approach fails to differentiate affiliative humour from other styles, particularly aggressive humour, as shown in Figure 3, highlighting a critical issue.

This misclassification may stem from affiliative humour sometimes containing slightly aggressive components, as noted by Martin et al. (2003). For example: **JOKE:** ‘To be happy with a man, you must understand him a lot and love him a little. To be happy with a woman, you must love her a lot and not try to understand her at all’. (**LABEL:** True: ‘Affiliative’, Predicted: ‘Aggressive’)

This joke attempts to playfully highlight gender differences, aiming for camaraderie. However, its misclassification as aggressive likely stems from the presence of gender stereotypes that could be misconstrued as demeaning. This example illustrates how subtle nuances in tone, context, and intent can lead to misclassifications between affiliative and aggressive humour.

5.2 Two-Model Approach

To address the challenge of misclassifying affiliative humour as aggressive, we implemented a two-model approach, consisting of a four-class model and a binary-class model. The performance of these individual models is presented in Tables 6 and 7, which show their accuracy and macro-mean f1-score, respectively. Among the four-class models, MUL+XGBoost achieved the highest performance, with an accuracy of 85.3% and a macro-mean f1-score of 85.1%. In contrast, the binary-class model ALI+XGBoost outperformed the other models, with an accuracy and f1-score of 80.0%.

The results of the two-model approach, which combines the four-class and binary models, are presented in Tables 8 and 9. This approach yields improved overall performance compared to the single-model method, with the best results

Model	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Five-Class	62.5	69.2	68.5	69.7	67.0	70.4	71.9	69.7	71.2	72.1	71.3	73.0	76.1	75.9
Four-Class	66.0	75.4	74.3	74.3	72.8	76.5	79.1	78.3	78.2	79.8	79.1	78.8	82.1	82.4
Binary-Class	74.8	73.9	78.8	75.9	74.8	77.2	78.1	71.9	79.5	74.1	75.2	76.6	80.3	78.3

Table 2: Mean Accuracy of 5-Fold Cross-Validation for the Various Classification Models

Model	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Five-Class	61.4	65.2	65.9	65.9	63.5	67.9	69.0	67.7	70.1	71.0	70.1	71.6	74.9	74.6
Four-Class	63.7	73.1	72.5	73.5	71.66	75.3	77.9	77.2	77.9	79.6	78.7	78.2	82.0	81.9
Binary-Class	74.1	71.4	77.1	73.8	72.9	75.6	76.1	70.0	78.8	73.1	73.8	75.4	79.5	77.7

Table 3: Macro-mean F1-Score of 5-Fold Cross-Validation for the Various Classification Models

	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Precision	64.1	72.7	71.4	76.5	65.0	72.9	72.7	73.6	70.1	73.7	68.5	77.6	76.8	75.6
Recall	62.5	70.3	71.6	74.3	64.0	72.7	72.1	74.0	70.6	72.7	68.4	77.6	77.4	75.1
F1-score	61.4	68.5	69.2	72.6	61.7	71.8	70.8	72.6	69.7	72.3	67.6	77.3	76.6	75.2
Accuracy	61.8	70.3	71.7	74.4	64.5	73.0	72.7	73.7	71.3	73.0	68.9	77.8	77.1	75.4

Table 4: Performance of the Single-Model Approach

Humour Styles	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Self-enhancing	61.7	80.3	81.9	82.8	70.7	76.9	85.0	80.3	81.6	80.0	73.2	82.6	86.2	79.4
Self-deprecating	66.0	72.5	76.7	80.5	65.9	70.5	66.7	77.1	67.4	75.9	71.3	79.1	77.6	76.7
Affiliative	39.2	40.5	34.9	46.5	33.7	54.5	47.3	50.0	48.5	57.4	48.0	64.9	63.0	60.2
Aggressive	56.4	62.7	69.1	72.0	58.9	71.3	67.6	67.2	65.6	66.7	62.8	74.8	67.7	70.8
Neutral	83.6	86.3	83.4	81.3	79.5	85.7	87.1	88.2	85.3	81.6	82.6	85.1	88.7	88.7

Table 5: Macro-mean F1-score for each humour style for the Single-Model Approach

achieved by the combination of MUL+XGBoost and ALI+XGBoost, which attained a f1-score of 78.0% and an accuracy of 77.8%. Notably, in Tables 8 and 9, MUL+XGBoost was consistently used as the four-class model in combination with various binary models (embeddings + RF or XGBoost), as it had previously demonstrated the best performance among the four-class models.

The Wilcoxon signed-rank test results (Table 10) statistically validate the improvements observed in the two-model approach. Significant improvements (p-value < 0.05) are evident for most metrics and humour styles, except aggressive humour (p-value = 0.1189). The two-model approach consistently outperforms the single-model approach, with average increases ranging from 3.42% to 4.91% across precision, recall, f1-score, and accuracy.

Notably, the two-model approach significantly improved affiliative humour classification, with an 11.61% increase in f1-score. All 14 models showed improvement for affiliative humour under this approach, suggesting more robust and accurate classification, especially for previously challenging categories like affiliative humour.

The cross-validation results (Tables 2 and 3) further support the robustness of our findings. The five-class models’ cross-validation accuracies and macro-mean f1-scores generally align with final test set accuracies and macro-mean f1-scores, indicating good generalisation. The four-class and binary-class models achieved even closer alignment, suggesting robust generalisation.

In summary, the two-model approach demonstrates superior performance in humour style recognition, particularly in identifying affiliative humour, with improved performance and generalisability across various metrics.

6 Conclusion

Automatic recognition of humour styles is a valuable yet challenging task with significant implications for digital humanities research, particularly in areas such as mental health, content moderation, and social media discourse. This study addresses the lack of established resources by introducing a new dataset of 1,463 instances across four humour styles and non-humour, while providing baseline evaluations of various models.

	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
Models		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Four-Class	73.0	76.5	80.5	77.1	75.1	80.9	83.6	80.5	80.9	81.2	76.8	82.6	85.3	82.6
Binary-Class	76.7	70.0	74.2	74.2	73.3	75.8	78.3	71.7	71.7	74.2	70.8	80.0	78.3	79.2

Table 6: Performance Accuracy of Four-Class and Binary-Class Individual Models

	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
Models		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Four-Class	70.5	73.2	78.4	76.1	73.5	80.1	82.4	79.4	78.8	80.3	75.7	81.3	85.1	81.8
Binary-Class	76.3	69.5	73.9	73.8	73.1	75.8	78.3	71.6	71.4	73.9	70.7	80.0	78.3	79.2

Table 7: Macro-mean F1-score of Four-Class and Binary-Class Individual Models

Four-Class Model ->	NB	MUL + XGBoost						MUL + XGBoost						DistilBERT
Binary-Class Model ->	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Precision	72.7	75.0	77.9	78.1	78.3	78.6	78.5	75.5	76.3	77.2	76.5	78.6	78.2	76.8
Recall	67.2	73.6	76.3	76.3	76.3	76.9	77.2	74.5	74.9	75.9	74.9	77.8	77.5	74.8
F1-score	67.4	73.5	76.3	76.2	76.3	77.1	77.4	74.8	75.0	75.9	75.1	78.0	77.7	75.3
Accuracy	67.6	73.4	76.1	76.1	76.1	76.8	77.1	74.4	74.7	75.8	74.7	77.8	77.5	75.4

Table 8: Performance of the Two-Model Approach

Four-Class Model ->	NB	MUL + XGBoost						MUL + XGBoost						DistilBERT
Binary-Class Model ->	NB (%)	Random Forest (%)						XGBoost (%)						DistilBERT (%)
		BGE	GTE	UAE	MRL	ALI	MUL	BGE	GTE	UAE	MRL	ALI	MUL	
Self-enhancing	56.8	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	80.3
Self-deprecating	66.7	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	75.6
Affiliative	67.6	50.5	59.8	59.0	61.0	66.7	65.5	57.1	56.4	58.2	58.2	66.1	63.9	61.2
Aggressive	64.0	61.5	65.7	66.2	64.8	63.3	65.7	61.1	62.9	65.7	61.4	68.2	68.8	71.2
Neutral	81.7	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.1

Table 9: Macro-mean F1-score for each humour style for the Two-Model Approach

	Precision	Recall	F1-Score	Accuracy	Self-enhancing	Self-deprecating	Affiliative	Aggressive	Neutral
Wilcoxon Statistics	0.0	3.0	0.0	0.0	8.0	3.0	0.0	27.0	10.0
P-value	0.000122	0.000610	0.000122	0.00220	0.0031	0.0006	0.0001	0.1189	0.0052
Average (Single-Model)	72.23	71.66	70.52	71.83	78.76	73.14	49.19	66.69	84.79
Average (Two-Model)	77.01	75.29	75.43	75.25	83.85	79.51	60.80	65.04	87.90
Model Difference	4.79	3.63	4.91	3.42	5.09	6.37	11.61	-1.65	3.11
# of improved models out of 14	14	13	14	12	13	13	14	4	11

Table 10: Wilcoxon Sign-Rank Test to Compare the Single-model and Two-model Approaches

The dataset and research have significant implications in three key areas:

1. **Mental Health:** Automatically identifying humour styles can enhance mental health research by enabling large-scale analysis of social media content. Different humour styles may correlate with various mental health indicators, potentially aiding in early detection of conditions such as depression or anxiety. For example, frequent use of self-deprecating humour might signal underlying mental health concerns.
2. **Content Moderation:** The dataset can contribute to more refined content moderation systems on social media platforms. By distinguishing between different humour styles, moderators can better identify potentially harmful content disguised as humour, such

as aggressive or self-defeating jokes, while allowing for benign forms of humour that enhance online interactions.

3. **Social Media Discourse:** Automatic recognition of humour styles can provide valuable insights into social dynamics and communication patterns across various online communities. This can help researchers understand how different humour styles influence online discussions, shape public opinion, and contribute to the spread of information or misinformation.

Our initial single-model approach struggled to accurately recognise affiliative humour, with f1-scores ranging from 39.2% to 64.9%. To address this, we developed a two-model approach consisting of a four-class model (merging affiliative and aggressive styles) followed by a binary model dis-

tinguishing between these styles. Extensive evaluation demonstrated the effectiveness of this approach in improving affiliative humour recognition, achieving f1-scores of 50.5% to 66.1%, while maintaining good performance for other styles. Furthermore, this approach offers flexibility in combining the best models for each sub-task, optimising overall performance.

By introducing this dataset and baseline evaluations, we aim to catalyse further research and development in these critical areas of digital humanities, ultimately enhancing our understanding of humour and its multifaceted impact on human communication.

7 Dataset Availability

The dataset and models implemented in this study are available to the community via the link in the footnote². Additionally, thirty instances from the dataset are included in Appendix E.

8 Limitations and Future Works

This study has several limitations. The dataset, consisting of 1,463 instances, is relatively small, which may limit the model's generalisation capabilities. Additionally, the inherent subjectivity of humour, along with the observed inter-rater agreement and annotation disagreements, underscores the challenges in consistently labelling humorous content. The focus on English-centric jokes may also introduce biases and language-specific nuances.

Future research could focus on collecting larger and more diverse datasets from various languages and sources to improve the robustness of the model. Leveraging transfer learning methods, such as intermediate fine-tuning on pre-trained language models, could enhance performance, especially when data is limited. Exploring multimodal approaches that incorporate visual, auditory, and contextual cues, as well as personalised models that adapt to individual preferences, could provide deeper insights into humour styles. Furthermore, investigating generative models for producing humorous content in specific styles presents a promising direction for further exploration.

Despite these limitations, this study lays the groundwork for humour style recognition, paving the way for extensive future research on computational humour analysis and its applications in

digital humanities.

9 Acknowledgement

This research was supported by the Petroleum Technology Development Fund (PTDF) of Nigeria.

References

- Arooba Amjad and Rabia Dasti. 2022. *Humor styles, emotion regulation and subjective well-being in young adults*. *Current Psychology*, 41(9):6326–6335.
- Issa Annamoradnejad and Gohar Zoghi. 2020. *ColBERT: Using BERT Sentence Embedding in Parallel Neural Networks for Computational Humor*.
- Berfu B, Ali H, Arzucan "ozgür Arzucan", and Arzucan "ozg Arzucan" "ozg" Arzucan "ozgür. 2020. *Analyzing ELMo and DistilBERT on Socio-political News Classification*. In *Language Resources and Evaluation Conference*, pages 11–16.
- Daniel Berrar. 2019. *Bayes' Theorem and Naive Bayes Classifier*. *Encyclopedia of Bioinformatics and Computational Biology*, (January 2018):0–18.
- Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. *All-in-One: A Deep Attentive Multi-task Learning Framework for Humour, Sarcasm, Offensive, Motivation, and Sentiment on Memes*. Technical report.
- Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2022a. *The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress*. In *MuSe 2022 - Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge*, pages 5–14. Association for Computing Machinery, Inc.
- Lukas Christ, Shahin Amiriparian, Alexander Kathan, Niklas Müller, Andreas König, and Björn W. Schuller. 2022b. *Multimodal Prediction of Spontaneous Humour: A Novel Dataset and First Results*. *Transactions on Affective Computing*, 20(10):1–15.
- Cleveland Clinic. 2024. *Self-Love: Why It's Important and What You Can Do To Love Yourself*.
- Silvie Cooper and David Dickinson. 2013. *Just jokes! Icebreakers, innuendo, teasing and talking: The role of humour in HIV/AIDS peer education among university students*. *African Journal of AIDS Research*, 12(4):229–238.
- Ron Deiter. 2000. *The Use of Humor as a Teaching Tool in the College Classroom*. Technical Report 2.

²Humour Styles Dataset: https://github.com/MaryKenneth/Two_Model_Humour_Style

- Liliya A. Demidova. 2021. [Two-stage hybrid data classifiers based on svm and knn algorithms](#). *Symmetry*, 13(4).
- Diwakar and Deepa Raj. 2024. [DistilBERT-based Text Classification for Automated Diagnosis of Mental Health Conditions](#). In *Microbial Data Intelligence and Computational Techniques for Sustainable Computing*, pages 93–106.
- Abbas Edalat. 2023. [Self-initiated humour protocols: An algorithmic approach for learning to laugh](#). *PsyArXiv*, 5:1–14.
- Hong Fang, Dahao Liang, and Weiyu Xiang. 2024. [Single-Stage Extensive Semantic Fusion for multi-modal sarcasm detection](#). *Array*, 22:100344.
- William Kodom Gyasi. 2023. [Humor as an ice breaker in marital tension: A family communication perspective](#). *Mediterranean Journal of Social & Behavioral Research*, 7(2):103–111.
- William P Hampes. 2007. [The Relation Between Humor Styles and Empathy](#). *Europe's Journal of Psychology*, 6(3):34–45.
- He He, Nanyun Peng, and Percy Liang. 2019. [Pun Generation with Surprise](#). In *Proceedings of NAACL-HLT*, pages 1734–1744, Minneapolis. Association for Computational Linguistics.
- Lena Hedin, Ingrid Höjer, and Elinor Brunnberg. 2012. [Jokes and routines make everyday life a good life-on 'doing family' for young people in foster care in Sweden](#). *European Journal of Social Work*, 15(5):613–628.
- Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2022. [Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue](#). *Frontiers in Robotics and AI*, 9.
- Arunima Jaiswal, Monika Anshu, Mathur Prachi, and Sheena Mattu. 2019. [Automatic Humour Detection in Tweets using Soft Computing Paradigms](#). In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con)*, pages 172–176.
- Daniela Jeder. 2015. [Implications of Using Humor in the Classroom](#). *Procedia - Social and Behavioral Sciences*, 180:828–833.
- Yu Jiang, Guoxiang Tong, Henan Yin, and Naixue Xiong. 2019. [A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters](#). *IEEE Access*, 7:118310–118321.
- Wei Jin. 2020. [Research on Machine Learning and Its Algorithms and Development](#). *Journal of Physics: Conference Series*, 1544:1–6.
- Jarrad Jinks. 2023. [Intermediate Task Ensembling for Sarcasm Detection](#). In *Bramer, M., Stahl, F. (eds) Artificial Intelligence XL*, volume 14381, pages 19–32. Springer, Charm, Oldeenburg.
- Ashraf Kamal and Muhammad Abulaish. 2020. [Self-deprecating Humor Detection: A Machine Learning Approach](#). In Le-Minh Nguyen, Xuan-Hieu Phan, Kôiti Hasida, and Satoshi Tojo, editors, *Computer Linguistics*, volume 1215 of *Communications in Computer and Information Science*, pages 483–484. Springer Singapore, Singapore.
- Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024. [Systematic Literature Review: Computational Approaches for Humour Style Classification](#). Technical report.
- Muhammad Umar Khan, Sumair Aziz, Khushbakht Iqtidar, Galila Faisal Zaher, Shareefa Alghamdi, and Munazza Gull. 2022. [A two-stage classification model integrating feature fusion for coronary artery disease detection and classification](#). *Multimedia Tools and Applications*, 81(10):13661–13690.
- I. I. Khramtsova and T. S. Chuykova. 2016. [Mindfulness and self-compassion as predictors of humor styles in US and Russia](#). *Social Psychology and Society*, 7(2):93–108.
- Nicholas Kuiper, Gillian Kirsh, and Nadia Maiolino. 2016. [Identity and Intimacy Development, Humor Styles, and Psychological Well-Being](#). *Identity*, 16(2):115–125.
- Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. [Open Source Strikes Bread - New Fluffy Embeddings Model \(emb2024mxbai\)](#).
- Xianming Li and Jing Li. 2023. [Angle-optimized Text Embeddings](#). *arXiv preprint*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards General Text Embeddings with Multi-stage Contrastive Learning](#). *arXiv preprint*.
- Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. [Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs](#). In *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, pages 4707–4715. Association for Computing Machinery, Inc.
- Jackson G. Lu. 2023. [Cultural differences in humor: A systematic review and critique](#).
- Fuli Luo, Shun Yao Li, Pengcheng Yang, Lei li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. [Pun-GAN: Generative Adversarial Network for Pun Generation](#).
- Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*, 2nd edition. Academic Press.
- Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. [Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire](#). *Journal of Research in Personality*, 37:48–75.

- Takuto Matsuda and Yoshiko Arimoto. 2023. [Detection of laughter and screaming using the attention and CTC models](#). In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023-August, pages 1025–1029. International Speech Communication Association.
- John Morreall. 2011. *Comic relief: A comprehensive philosophy of humor*. John Wiley & Sons, Ltd.
- John Morreall. 2012. [Philosophy of Humor](#).
- National Institute of Mental Health. 2024. [Caring for Your Mental Health](#).
- Hugo Gonalo Oliveira, Andr e Clem ncio, and Ana Alves. 2020. [Corpora and Baselines for Humour Recognition in Portuguese](#). In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1278–1285, Marseille. European Language Resources Association.
- Constantin Y. Plessen, Fabian R. Franken, Christoph Ster, Rebecca R. Schmid, Christoph Wolfmayr, Anna Maria Mayer, Marc Sobisch, Maximilian Kathofer, Katrin Rattner, Elona Kotlyar, Rory J. Maiерwieser, and Ulrich S. Tran. 2020. [Humor styles and personality: A systematic review and meta-analysis on the relations between humor styles and the Big Five personality traits](#). *Personality and Individual Differences*, 154.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS*.
- Tabea Scheel and Christine Gockel. 2017. *Humor at Work in Teams, Leadership, Negotiations, Learning and Health*, volume 1.
- Amir Schur and Sam Groenjes. 2024. [Comparative Analysis for Open-Source Large Language Models](#). In *Communications in Computer and Information Science*, volume 1958 CCIS, pages 48–54. Springer Science and Business Media Deutschland GmbH.
- Chen Shani, Nadav Borenstein, and Dafna Shahaf. 2021. [How Did This Get Funded?! Automatically Identifying Quirky Scientific Achievements](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 14–28. Association for Computational Linguistics.
- Carlo Strapparava, Oliviero Stock, and Rada Mihalcea. 2011. [Computational Humour](#). In *Emotion-Oriented systems*, pages 609–634.
- Hiroki Tanaka and Nick Campbell. 2014. [Classification of social laughter in natural conversational speech](#). *Computer Speech and Language*, 28(1):314–325.
- Leonard Tang, Alexander Cai, Steve Li, and Jason Wang. 2022. [The Naughtyformer: A Transformer Understands Offensive Humor](#). *arXiv*.
- Le Van Lam, Ian Welch, Xiaoying Gao, and Peter Komisarczuk. 2011. [Two-stage classification model to detect malicious web pages](#). In *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, pages 113–120.
- Jose Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. 2023. [Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild](#). *IEEE Transactions on Affective Computing*.
- Livia Veselka, Julie Aitken Schermer, Rod A. Martin, and Philip A. Vernon. 2010. [Relations between humor styles and the Dark Triad traits of personality](#). *Personality and Individual Differences*, 48(6):772–774.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual E5 Text Embeddings: A Technical Report](#). *arXiv preprint*.
- Orion Weller and Kevin Seppi. 2020. [The rJokes Dataset: a Large Scale Humor Collection](#). Technical report.
- Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. [RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder](#). *arXiv preprint*.
- Jong Eun Yim. 2016. [Therapeutic benefits of laughter in mental health: A theoretical review](#).
- Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. [A Neural Approach to Pun Generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 1650–1660. Association for Computational Linguistics.
- Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. [Retrieve Anything To Augment Large Language Models](#). *arXiv preprint*.

A Humour Style Websites

The website sources for the different humour styles, along with their corresponding links, are listed in Table 11.

Humour Styles	Website
Aggressive	Parade
	Laugh factory
	Reader's digest
	Pun Me
Affiliative	Reader's digest
	Independent
	Happy Numbers
	Laugh factory
	Team building
Self-Deprecating	Tastefully Offensive
	Bored Pandas
	Cracked
	Reddit
Self-Enhancing	Buzz Feed
	Put the Kettle On
	Silk and Sonder
	Carley Schweet
	Joyful through it all
	Laura Conteuse

Table 11: List of websites from which jokes were taken.

B Mapping Jokes to Humour Style Labels

Although certain humour websites from which the jokes were extracted do not explicitly categorise the humour as "aggressive," "affiliative," or "self-enhancing," there are reasonable justifications for associating the humour found on those sites with the respective humour styles, based on the content and intended audience. This section outlines the keywords and rationale for mapping jokes to the original labels, definitions, or tags provided for jokes on the websites.

B.1 Aggressive Humour

Aggressive humour is characterised by jokes, insults, or humorous remarks that are intended to disparage, belittle, or target particular individuals or groups. This type of humour often involves sarcasm, mockery, and put-downs, and it can be perceived as offensive or hostile by the targeted parties.

B.1.1 Equivalence classes (Website Keywords)

- **Dark (inappropriate) Jokes:** Dark (inappropriate) jokes are identified as being cruel, morbid, or offensive to some, which aligns with the characteristics of aggressive humour (Tang et al., 2022).
- **Insult:** Insult is an offensive remark or action intended to mock or belittle the target (Cambridge Dictionary (<https://rb.gy/10b2sz>)).

Insult is a key characteristic of aggressive humour (Martin et al., 2003).

B.2 Affiliative Humour

Affiliative humour is characterised by jokes, witty remarks, or humorous anecdotes that are intended to amuse others, facilitate social interactions, and strengthen relationships. This type of humour is non-hostile, benign, and often used to create a positive, inclusive atmosphere.

B.2.1 Equivalence classes (Website Keywords)

- **Icebreakers Jokes for Work Meetings:** These jokes are typically used to create a relaxed and friendly environment in professional or group settings. They are meant to facilitate social interactions and put people at ease, which aligns with the goals of affiliative humour (Cooper and Dickinson, 2013).
- **International Day of Happiness:** Jokes shared on occasions like the International Day of Happiness are typically intended to spread positivity, joy, and laughter among people. Such jokes are designed to bring people together and create a shared experience of amusement, which aligns with the goals of affiliative humour.
- **Friendship:** Jokes meant to be shared among friends are often used to strengthen bonds, create shared laughter experiences, and reinforce the positive aspects of friendship. This type of humour is non-threatening and aimed at building connections, which is a characteristic of affiliative humour.
- **Family jokes:** Jokes shared within families are often intended to create a sense of bonding, shared laughter, and enjoyment. Family jokes are generally non-offensive and serve to strengthen familial relationships, which is a characteristic of affiliative humour (Hedin et al., 2012; Gyasi, 2023).
- **Classroom:** Humour shared between teachers and students, or within educational settings, is often used to create a positive and engaging learning environment. These jokes are likely meant to connect with students and foster a sense of camaraderie, which is in line with affiliative humour (Deiter, 2000; Jeder, 2015).

B.3 Self-deprecating Humour

Self-deprecating humour is a type of humour in which individuals make fun of their own flaws,

to.

- **Aggressive (A2):** In contrast, one annotator interpreted the joke as aggressive, as the Devil’s representation of negative traits or behaviours could be seen as a critique of people in general.

E Sample Jokes Dataset

In this section, we showcase a random selection of thirty samples from our jokes dataset (see Table 13). Each sample consists of the joke content paired with its corresponding label, providing a glimpse into the dataset’s composition and structure. For reference, the labels are interpreted as follows:

- Self-enhancing: 0
- Self-deprecating: 1
- Affiliative: 2
- Aggressive: 3
- Neutral: 4

Jokes	A1	A2	A3	Hugging Chat	Claude	Copilot	Chat-GPT
Insanity is hereditary, - You get it from your children.	0	3	2	1	0	2	2
Gravity doesn't exist: the earth sucks.	0	4	3	4	0	4	4
Did you hear about the Scottish Kamikaze pilot? He crashed his plane in his brother's junkyard	3	4	2	3	3	3	3
Biology grows on you	4	3	2	2	4	4	4
Don't worry if you're a kleptomaniac, you can always take something for it	0	3	2	1	2	2	2
To steal from one is plagiarism. To steal from many is research	2	3	4	1	0	2	2
If all else fails, lower your standards	2	1	4	1	1	1	1
There are only 3 things that tell the truth: 1 - Young Children 2 - Drunks 3 - Leggings	1	3	4	2	4	2	2
Never get stuck behind the Devil in a Post Office queue! The Devil can take many forms.	2	3	4	4	2	4	4
Always follow your dreams. Except for that one where you're naked at work.	2	0	4	1	1	2	2

Table 12: Annotation Disagreement

Jokes	Labels
Is that your nose or are you eating a banana?	3
Q: Why did the witches' team lose the baseball game? A: Their bats flew away.	2
Act your age, not your shoe size.	3
I may be trash, but I burn with a bright flame	1
Yeah, I know. I hate me too.	1
"The secret of staying young is to live honestly, eat slowly, and lie about your age."	0
"I got it all together. But I forgot where I put it."	0
A man on a date wonders if he'll get lucky. A woman already knows.	2
Here's how unfair the tax system is in each state	4
Is a death sentence really a death sentence?	4
Trump's new military plan will cost 150 billion dollars – at the very least	4
He is so short, his hair smells like feet.	3
You should be in commercials for birth control.	3
"The road to success is dotted with many tempting parking spaces."	0
If I had a face like yours, I'd sue my parents!	3
"I'm not perfect, but I'm perfectly me."	0
Why don't scientists trust atoms? Because they make up everything!	2
Don't mind me. I'm just having an existential crisis. Move along, folks.	1
I can't talk to you right now, tell me, where will you be in 10 years?	3
A wise woman once said, "fuck this shit" and lived happily ever after.	0
He is depriving a village somewhere of an idiot.	3
Dad: "Can I see your report card, son?" Son: "I don't have it." Dad: "Why?" Son: "I gave it to my friend. He wanted to scare his parents."	2
"The elevator to success is out of order. You'll have to use the stairs, one step at a time."	0
"I'm not the kind of guy who has a huge weight problem, but I am the kind of guy who could really put the brakes on an orgy. Everyone would be like, 'Was he invited? Why is he eating a cake?' I've never been in an orgy, but I feel like it'd be like what happens when I try to play pickup basketball: No one passes me the ball, and everyone asks me to keep my shirt on."	1
"I'm a self-love junkie. Can't get enough of this good stuff!"	0
"If I could rearrange the alphabet, I'd put 'U' and 'I' together."	2
"Let your light shine bright so the other weirdos can't find you"	0
Did you hear about the magic tractor? It turned into a field.	2
I don't have a nervous system. I am a nervous system!	1
How to build muscle: proven strength lessons from milo of croton	4

Table 13: Samples from the Humour Styles Dataset

N-gram-Based Preprocessing for Sandhi Reversion in Vedic Sanskrit

Yuzuki Tsukagoshi and Ikki Ohmukai

The University of Tokyo / Tokyo, Japan

{yuzuki, i2k}@l.u-tokyo.ac.jp

Abstract

This study aims to address the challenges posed by sandhi in Vedic Sanskrit, a phenomenon that complicates the computational analysis of Sanskrit texts. By focusing on sandhi reversion, the research seeks to improve the accuracy of processing Vedic Sanskrit, an older layer of the language. Sandhi, a phonological phenomenon, poses challenges for text processing in Sanskrit due to the fusion of word boundaries or the sound change around word boundaries. In this research, we developed a transformer-based model with a novel n-gram preprocessing strategy to improve the accuracy of sandhi reversion for Vedic. We created character-based n-gram texts of varying lengths ($n = 2, 3, 4, 5, 6$) from the Rigveda, the oldest Vedic text, and trained models on these texts to perform machine translation from post-sandhi to pre-sandhi forms. In the results, we found that the model trained with 5-gram text achieved the highest accuracy. This success is likely due to the 5-gram's ability to capture the maximum phonemic context in which Vedic sandhi occurs, making it more effective for the task. These findings suggest that by leveraging the inherent characteristics of phonological changes in language, even simple preprocessing methods like n-gram segmentation can significantly improve the accuracy of complex linguistic tasks.

1 Introduction

1.1 Sandhi in Sanskrit

Vedic Sanskrit represents an earlier stage of the Sanskrit language, characterized notably by its unique system of accentuation, where the meaning of words in Vedic can vary depending on the position of the accent.

Sandhi refers to a phonological/phonetic phenomenon wherein certain conditions induce changes in the sounds of adjacent words or morphemes (Macdonell, 1910). This phenomenon

is divided into two types: external sandhi, which involves sound changes at the junctions between words in a sequence, and internal sandhi, which affects the internal structure of words, such as in the derivation of nouns and verbs.

For instance, when the words *yás* (“who/which”, a relative pronoun in the nominative singular masculine), *hatvá* (“having killed”, absolutive of the verb *han*), and *áhim* (“a snake”, a noun in the accusative singular) consecutively in a sentence, they have a surface form like *yó hatváhim* (Rigveda 2.12.3a) meaning “[Indra,] who [...] after killing the snake”. The final two sounds of *yás* change to *-ó* before the voiced sound *h-*. Additionally, the final vowel *-á* of *hatvá* and the initial vowel *á-* of *áhim* merge into *-á-*. Conversely, when the same word *hatvá* is followed by *pr̥thivyám* (“on the earth”), it becomes *hatvá pr̥thivyám* (Rigveda 1.100.18b), meaning “After [he] struck [Dasyus and Śimyus, felled them] on the earth.” In this case, the final vowel *-á* of *hatvá* does not alter before the voiceless stop *p-*. Hence, external sandhi varies according to the boundary sounds of adjacent words in a sentence. Internal sandhi, in contrast, can alter the sound at the morpheme boundary, as illustrated by the example of the genitive singular form of the stem *tanú-* (“body”) is *tanv-ás*. Internal sandhi also undergoes variations depending on the adjacent sounds, similar to external sandhi.

In general, Sanskrit texts preserve the forms resulting from sandhi rules as they are. Therefore, to analyze Sanskrit texts, it is essential to identify the original word forms prior to the sound changes imposed by sandhi.

Despite the precise definition of sandhi rules, reversing the transformations to recover pre-sandhi forms poses significant challenges, whether done manually or via computational methods. To accurately revert *hatváhim* back to the two separate words *hatvá* and *áhim*, one must have knowledge of the word forms: the absolutive *hatvá* of the verb

han (“kill”) and the accusative singular *áhim* of the noun *áhi-* (“snake”). When attempting to segment the phonetic sequence h-a-t-v-á-h-i-m into two parts, one must first determine the necessity of splitting. If splitting is required, a sequence containing a long vowel *-ā-* without consideration of accent, may yield four potential outcomes: *-ā ā-*, *-ā ā-*, *-ā ā-*, and *-ā ā-* (For clarity, short vowels are indicated with *ă*). Furthermore, when taking accent into account, numerous additional possibilities arise, including sequences where both parts bear an accent, or sequences where only one part is accented, e.g. *ihásti = ihá asti, índrâ = indra á*.

1.2 Necessity of Sandhi Reversion

The process of restoring word sequences in a text from their post-sandhi forms back to their original, pre-sandhi forms is termed *sandhi reversion* or the act of *reverting sandhi* in this paper. Sandhi reversion is essential for accurately understanding and analyzing Sanskrit texts, as the phenomenon of sandhi causes phonetic changes that can obscure the original word boundaries. To perform sandhi reversion effectively, it is necessary to consider not only the given phonetic sequence but also the morphological and syntactic context in which these forms appear.

As discussed in the previous section (section 1.1), Sanskrit texts are predominantly written in their post-sandhi forms. This convention presents a challenge for linguistic analysis and computational processing, as it complicates the direct extraction and identification of individual words from the text. Thus, sandhi reversion becomes a crucial preprocessing step in many aspects of Sanskrit language processing, enabling scholars and computational tools to work with the underlying word forms rather than their modified, surface-level representations.

The reversion of external sandhi, in particular, which involves converting a continuous sentence into a series of discrete words, is vital in various domains such as natural language processing (NLP), philology, and linguistics. External sandhi impacts word recognition and syntactic analysis, and its accurate reversion is necessary to ensure proper word segmentation and syntactic parsing in these fields. Therefore, this study focuses on external sandhi. For the remainder of this paper, the term “sandhi” will refer exclusively to external sandhi, reflecting its primary importance for our objectives.

2 Related Work

Given the vast corpus of Sanskrit texts, manually reverting all instances of sandhi is impractical and time-consuming. As a result, numerous computational methods for sandhi reversion have been developed over the years to automate this process.

Recent research has shown that machine learning-based sandhi reversion is more accurate than previous rule-based approaches. The following sections review major studies that have employed machine learning techniques for sandhi reversion.

2.1 seq2seq + Attention / Classic

Reddy et al. (2018) approached sandhi reversion by treating a post-sandhi sentence as the source text and a pre-sandhi sentence as the target text, effectively framing the problem as a translation task. They applied a sequence-to-sequence (seq2seq) model with attention mechanisms, utilizing SentencePiece (Schuster and Nakajima, 2012) for preprocessing. This approach facilitated the creation of a translation model that could convert post-sandhi sentences back to their pre-sandhi forms. The study utilized a dataset (Krishna et al., 2017) of approximately 100,000 sentences to train the model.

The method involved segmenting sentences as demonstrated below, where spaces (represented by underscores, `_`) are treated as individual characters, and each resulting segment is considered a word within the restructured sentence:

(original) putram̐ vaṁśakaram̐ rāma ṅṛ-
pasamnidhau

(sentencepiece) _putram̐ _vaṁś akar
am̐_rāma ṅṛpa [sic.]¹ samnidh au

2.2 RNN + CNN / Classic + Vedic

Hellwig and Nehrdich (2018) developed a neural network model that combines Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) using character-based n-grams to address the challenge of sandhi reversion in Sanskrit texts. Their study evaluated the performance of various models, including the seq2seq approach by Reddy et al. (2018), and demonstrated that their CNN+RNN model outperformed previous methods in terms of accuracy and effectiveness for sandhi reversion.

¹An underscore should be placed before *ṅṛpa*.

The dataset used in the study comprised both Classical Sanskrit and Vedic Sanskrit texts. However, it is important to note that the Vedic Sanskrit texts in the dataset had their accent marks removed. As a result, the sandhi reversion model developed in this study cannot be directly applied to Vedic Sanskrit texts that include accent marks. In Vedic Sanskrit, where the meaning can change based on the position of the accent within a word, the absence of accent information in the dataset prevents accurate reconstruction of the original accentuation patterns. Thus, restoring the accents in Vedic Sanskrit from texts without them is not feasible using this model.

3 Method

This study addresses the task of translating text from post-sandhi to pre-sandhi forms, building upon the approaches discussed in the related work section (section 2). The objective is to develop a model that accurately reverses the sandhi processes, thereby restoring the original word boundaries in Vedic Sanskrit texts.

3.1 N-gram translation

In this study, a new “sentence” is treated as one that is composed of character-based n-grams of a given sentence. A character-based n-gram divides the sentence into n-character units.

For example, $y/o+hatv/\bar{a}him$ (= $y\acute{o} hatv\bar{a}him$) shown in the section 1.1. When $n=3$, for both post-sandhi $y/o+hatv/\bar{a}him$ and pre-sandhi $y/a\grave{h} hatv/\bar{a}/ahim$ the texts result in the following n-grams:

$y/o /o+ o+h +ha hat atv tv/ v/\bar{a} /\bar{a}h \bar{a}hi$
him (post-sandhi)

$y/a /a\grave{h} a\grave{h}+ \grave{h}+h +ha hat atv tv/ v/\bar{a} /a+ \bar{a}+ /+a /ah ahi him$ (pre-sandhi)

In these examples, acute accent marks are represented by slashes, and spaces by plus signs.

This method differs from the n-gram method of Kitagawa and Komachi (2018) and its derivative by Hellwig and Nehrlich (2018). In their methods, each character-based input layer simultaneously embeds n-gram strings starting/ending with that character. Consequently, all n-grams within a specified range are used as input for a single model. In contrast, our method treats each segment arranged in n-grams as a word unit.

Character-based n-gram texts were generated for $n = 2, 3, 4, 5, 6$. In addition to these n-gram

variations, an original text was prepared with no further processing beyond transliteration. These six types of texts were utilized as training data, as described in section 3.2. The model, detailed in Section 3.3, was then trained on each set of texts to perform machine translation from post-sandhi to pre-sandhi forms. This comprehensive approach allows for the assessment of how different n-gram lengths affect the model’s ability to accurately perform sandhi reversion.

3.2 Training datasets

The primary text used in this study is the electronic version of the Rigveda provided by Martínez García and Gippert (1995). The Rigveda, a central text within the Vedic corpus, exists in several versions, including the Samhitāpāṭha, which represents the post-sandhi form, and the Padapāṭha, which provides a pre-sandhi form. The Padapāṭha is a later interpretative text based on the Samhitāpāṭha, created to disambiguate the text by resolving the effects of sandhi.

Due to the inherent ambiguities involved in sandhi reversion, the Padapāṭha is considered one of several possible interpretations of the Ṛgveda Samhitā. Despite this, having paired post-sandhi and pre-sandhi texts is invaluable for supervised machine learning tasks. With approximately 10,000 verses, the Rigveda provides a substantial and well-suited dataset for training models to perform sandhi reversion.

3.3 Model

In this study, we address the problem of sandhi reversion as a translation task, converting post-sandhi text (Samhitāpāṭha) into pre-sandhi text (Padapāṭha). To accomplish this, we developed a transformer-based translation model.

Our approach focuses on preprocessing and the preprocessing strategy is designed to be independent of the specific machine translation architecture, making it compatible with the state-of-the-art architecture.²

4 Results

Table 1 presents the Precision, Recall, and F1 scores for models with six different types of text: one model trained on the original, unprocessed text (denoted as “Word”) and five models trained on

²The model, dataset, and associated scripts is available at <https://github.com/Yuzki/SktTool>.

character-based n-gram texts with varying length ($n = 2, 3, 4, 5, 6$). These metrics provide a comprehensive evaluation of the model’s performance in translating post-sandhi to pre-sandhi forms across different preprocessing strategies.

For comparative purposes, Table 2 displays the performance metrics reported by Hellwig and Nehrdich (2018) for their combined CNN and RNN model, alongside results from earlier studies.

	Precision	Recall	F1
Word	95.0	96.6	95.8
2-gram	87.8	88.3	88.1
3-gram	94.8	95.9	95.4
4-gram	95.8	97.2	96.5
5-gram	96.0	97.7	96.8
6-gram	95.3	97.2	96.2

Table 1: Precision, Recall, and F1 of each n-gram text and the original text.

	Precision	Recall	F1
Hellwig (2015b)	91.8	91.8	94.8
Reddy et al. (2018)	90.2	88.4	93.3
Transformer 5K	94.9	94.5	96.5
rcNN _{short} ^{split}	94.6	94.8	96.7

Table 2: Excerpt from Table 3 in Hellwig and Nehrdich (2018).

The training and evaluation times for each model are not a primary focus of this study, because the computational efficiency is less of a concern for Sanskrit, an ancient language with a relatively fixed corpus, compared to the modern languages with increasing linguistic resources.

The results indicate that the sandhi reversion model trained on 5-gram text achieves the highest Precision, Recall, and F1 scores among all the models tested. Notably, these scores are slightly higher than those reported for the most accurate sandhi reversion model using the combined RNN + CNN approach by Hellwig and Nehrdich (2018), even though the datasets used in the two studies differ.

5 Limitations

While the proposed n-gram-based preprocessing method showed promising results, several limitations remain. First, the dataset used in this study

was limited to the Rigveda, and the model’s effectiveness on other Vedic or Classical Sanskrit texts remains to be evaluated. Additionally, although the 5-gram model performed well, this approach may not generalize to texts with different phonological structures, such as accentless Sanskrit texts. Another limitation is the reliance on the Padapāṭha as the pre-sandhi form, which is one of several possible interpretations of the Rigveda.

6 Conclusion

The results of this study demonstrate that some models utilizing n-gram preprocessing outperform the model with no special text processing. The highest accuracy is particularly achieved when $n = 5$ (Table 1).

This n value is thought to be related to the phonological environment in which sandhi occurs. The maximum number of phonemes in the environment where sandhi occurs is five: $/a\ s + X -;$ where $/$ denotes an accent, $+$ denotes a space, and X represents an arbitrary voiced sound. Therefore, when $n = 5$, the n-gram text adequately includes this sequence, potentially resulting in the highest accuracy. However, it is predicted that $n = 4$ will yield the highest accuracy for texts in Classical Sanskrit, which does not have accents.

By demonstrating the effectiveness of n-gram preprocessing, this study shows that the approach is independent of the specific translation architecture used. This flexibility implies that our method can be adapted to future innovations in translation tasks, regardless of the underlying architecture.

Although the corpus of Vedic texts is smaller than that of Classical texts, Vedic language processing remains in its early stages. Implementing sandhi reversion as a preprocessing technique significantly advances Vedic NLP by facilitating tasks such as named entity recognition and intertextual comparison based on word occurrences.

Furthermore, this n-gram segmentation method is not restricted to the sandhi reversion problem in Vedic Sanskrit; it is also applicable to word segmentation challenges in Classical Sanskrit, which lacks accents, as well as in other languages. In these contexts, using n-grams with a size close to or matching the maximum string length of the relevant phonetic sequences in each language is expected to yield high accuracy.

References

- Oliver Hellwig and Sebastian Nehrlich. 2018. [Sanskrit word segmentation using character-level recurrent and convolutional neural networks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763. Association for Computational Linguistics.
- Yoshiaki Kitagawa and Mamoru Komachi. 2018. [Long short-term memory for japanese word segmentation](#). In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics.
- Amrith Krishna, Pavan Kumar Satuluri, and Pawan Goyal. 2017. [A dataset for sanskrit word segmentation](#). In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 105–114. Association for Computational Linguistics.
- Arthur Anthony Macdonell. 1910. *Vedic grammar*. Verlag von Karl J. Trübner.
- F. J. Martínez García and J. Gippert. 1995. [Plain text retrieval, thesaurus indogermanischer text- und sprachmaterialien](#). Online; accessed 17 October 2022.
- Vikas Reddy, Amrith Krishna, Vishnu Sharma, Prateek Gupta, Vineeth M R, and Pawan Goyal. 2018. [Building a word segmenter for sanskrit overnight](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).
- Mike Schuster and Kaisuke Nakajima. 2012. [Japanese and korean voice search](#). In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Enhancing Swedish Parliamentary Data: Annotation, Accessibility, and Application in Digital Humanities

Shafqat Mumtaz Virk

University of Gothenburg, Sweden
shafqat.virk@svenska.gu.se

Claes Ohlsson

Linnaeus University, Sweden
claes.ohlsson@lnu.se

Henrik Björck

University of Gothenburg, Sweden
henrik.bjorck@lir.gu.se

Nina Tahmasebi

University of Gothenburg, Sweden
nina.tahmasebi@gu.se

Leif Runefelt

Södertörn University, Sweden
leif.runefelt@sh.se

Abstract

The Swedish bicameral parliament data presents a valuable textual resource that is of interest for many researchers and scholars. The parliamentary texts offer many avenues for research including the study of how various affairs were run by governments over time. The Parliament proceedings are available in textual format, but in their original form, they are noisy and unstructured and thus hard to explore and investigate. In this paper, we report the transformation of the raw bicameral parliament data (1867-1970) into a structured lexical resource annotated with various word and document level attributes. The annotated data is then made searchable through two modern corpus infrastructure components which provide a wide array of corpus exploration, visualization, and comparison options. To demonstrate the practical utility of this resource, we present a case study examining the transformation of the concept of 'market' over time from a tangible physical entity to an abstract idea.

1 Introduction

In recent years, the digitization of historical and contemporary text has facilitated valuable research in text-based fields particularly in digital humanities. While newspapers and literature offer important avenues, parliamentary text are both complementary and important sources of knowledge. In this paper, we present the digitized Swedish bicameral parliamentary data annotated with various token and text level attributes. The corpus is made accessible through the modern corpus infrastructure of Språkbanken, the Swedish Language Bank¹. By applying a range of token and text-level annotations, we aim to enhance the accessibility and usability of the parliamentary records, making them a valuable resource for a broad range of research inquiries.

¹<https://spraakbanken.gu.se/>

The annotation process involves several layers, from basic tokenization to more advanced linguistic tagging. For this purpose, we rely on several in-house developed and external annotation tools. These annotations not only help in structuring the data but also enable researchers to extract meaningful patterns and insights that would otherwise remain obscured. Furthermore, by integrating these annotated records into a modern and well-established corpus infrastructure, we ensure that the data is both easily accessible and scalable for various computational analyses.

To illustrate the practical value of our enriched parliamentary data, we present a case study that explores the evolution of market-related language within legislative and political discourse. Today, the market is a ubiquitous concept in both professional and private matters. However, this was not always the case historically, and 'the market' has gradually developed into the central concept it is today. While conceptual history studies have examined the market, particularly in later periods (Leary, 2019), a continuous description of the market as a concept across various contexts and discourses remains lacking. In response to this gap, a project called "The Market Language"² aims to provide an empirical based analysis of the market concept and its usage in Sweden, utilizing available corpora (Ohlsson et al., 2022).

Historically, the concept of "the market" has undergone transformation from an original meaning as a tangible, physical and time-specific space for trade, to more abstract notions like the "iron market," and eventually to a role of an active agent with the potential for influencing professional or private daily life, as in expressions such as "the market reacted badly to the latest inflation news." This linguistic and semantic shift reflects broader socio-economic changes and raises numerous research

²<https://www.gu.se/forskning/marknadens-sprak-studier-i-talet-om-marknader-fran-medeltid-till-nutid>

questions of interest to both linguists and historians and also scholars in several other disciplines. For instance, when and how did this conceptual shift in market language occur? Are there discernible linguistic patterns associated with different types of semantic change? Can linguistic changes in how “market” is used be related to social and historical development or changes and also vice versa?

The annotated parliamentary texts offer a rich and structured resource for investigating these questions. Through this case study, we demonstrate the potential of our annotated corpus not only as a tool for linguistic analysis but also as a gateway to deeper historical inquiries, as done in (De Bolla, 2023) as an example.

By enabling researchers to trace the usage, frequency, and context of market-related terms across time, our corpus can provide insights into the factors driving these linguistic transformations. Studying the change in meaning of the concept of “market” in the Swedish language offers several advantages as an example of how parliamentary data has been incorporated into the corpus infrastructure. The word “market,” or “marknad” in Swedish, has a consistent lexical form over time, making it a good candidate for annotation. Additionally, “market” is a concept that appears across various types of texts or discourses. This allows for comparisons between its use in political or legislative discourse and its use in other corpora within the infrastructure, such as media or literary texts. Furthermore, existing research and studies have shown examples and parts of the development of the word “market,” from the primarily concrete meaning that is still in use, to become increasingly abstract. The possibility to use a large set of discourse-specific data such as the parliamentary corpus, opens for showing a fuller picture of this development over time.

2 The Swedish Parliament Data

The history of the Swedish parliament or *Riksdag* starts in the 15th century when the Riksdag of the Estates was formed with roots in parliamentary gatherings of Swedish noblemen that span much longer. In the early modern era, this type of parliament was replaced with the bicameral parliament in 1867, which was in function until 1970 when the current one chamber parliament of the modern Riksdag was installed. The bicameral Riksdag existed at a time when Sweden as a society was subject to many changes and went from being a

poor, war-ridden and mostly rural country to becoming a modern and industrialized nation with a strong focus on democratic and emancipated values. The fundamentals of today’s democratic form of government were essentially laid out, debated, and tried during the periods of the late 1800s and first half of the 1900s when the bicameral parliament was in function. This makes the texts of the bicameral Riksdag important for researchers and scholars from several disciplines and there are other projects focusing on making other features than purely linguistic ones more available such as the The Open Parliament Laboratory (OPaL)³ at Örebro University in Sweden and also the SWERIK⁴ project. Similar efforts to make parliamentary data more available for especially research purposes can be found in the ParlaMint project that encompasses comparable corpora of parliamentary debates from 29 European countries and regions. This is a valuable resource for synchronic data, but the inclusion of the Swedish bicameral parliament data also makes older, historical datasets available for research.

The texts of the bicameral parliament of 1867 to 1970 were scanned by the Royal Library and are available in pdf format via the Riksdag website⁵ and the Royal Library of Sweden⁶. The downloadable data files contain some metadata regarding type of text and year but are otherwise unstructured and difficult to use for large-scale searches or for linguistic analysis purposes. The files are also presented per year, which makes the total sum of text files that are accessible relatively high. This makes the data set problematic to handle as a bundle for both researchers, and particularly, an interested public. The interface where the original pdf files can be reached is not developed for quick access to the material in full.

To overcome some of these obstacles, we have transformed the bicameral parliament data into a structured lexical resource by enriching it with a set of word and text level attributes. We have also made the data searchable through *Korp* and *Strix*, which are modern corpus infrastructure tools developed and maintained at Språkbanken Text⁷ (the Swedish Language Bank) at the University of

³<https://www.oru.se/english/research/research-environments/hs/opal--/>

⁴<https://swerik-project.github.io>

⁵<https://www.riksdagen.se/en/>

⁶<https://www.kb.se/in-english.html>

⁷<https://spraakbanken.gu.se/>

Gothenburg. These tools provide a range of options to easily access, explore, visualize, and compare the data that were not previously available.

3 The Corpus Categories and Statistics

The Swedish bicameral parliament data encompasses a variety of text types, including protocols, debates, and motions from individual members of parliament. The type categorization of the material is the original structure of the scanned text file and is used also for our integration of the parliamentary data into the corpus infrastructure, since the categories are based on the different text types of the parliamentary practices. The use of type categorization opens for more specialized searches and also for comparisons of different text types within the data set as well as with texts from both previous and contemporary parliamentary resources. Table 1 shows the list of corpus categories, and statistics on number of documents, sentences and tokens in each category. As can be noted, there are roughly 190k, 46M and 0.9B number of documents, sentences, and words respectively in total. The data contains some metadata as aforementioned (type and year information). These categories follow the rationale for parliamentary work over the period 1867 to 1970, with focus on the process of proposing and debating legislature. 10 different categories are distinguished (as shown in Table 1) and these are now searchable through the interfaces Korp and Strix.

4 The corpus infrastructure

In recent years, there has been a remarkable surge in the production of digital textual data, i.e., corpora, and the conversion of non-digital texts into digital formats. This has simultaneously driven the need for the development of efficient methods for storing and exploring these extensive datasets. Consequently, technology has evolved from basic string-matching search approaches to the creation of advanced corpus infrastructures that offer query-based search, comparison, and visualization capabilities. In the following sections, we will briefly introduce two such tools in the corpus infrastructure domain: Korp and Strix. These tools offer a wide array of options for exploring, comparing, and visualizing corpus and related statistics at word, sentence, and document levels.

4.1 Korp

Korp⁸ (Borin et al., 2012) is one of the key components of the corpus infrastructure developed and maintained by Språkbanken⁹ (the Swedish language bank). It comprises separate backend and frontend components designed for corpus storage and exploration. The backend is used for importing data into the infrastructure, annotating it, and converting it to various formats for downloading. Korp provides an annotation pipeline for adding a range of lexical, syntactic, and semantic annotations to the corpus, utilizing both internal and external annotation tools. On the other hand, the frontend offers a variety of search options, including basic, extended, and advanced search functionalities, enabling users to extract and visualize search results, annotations, statistics, and comparison between corpora. Section 5 provides some practical examples of these capabilities.

4.2 Strix

Strix¹⁰ is another tool within Språkbanken’s corpus infrastructure. While it shares similarities with Korp by offering search and exploration capabilities for text collections and their annotations, it distinguishes itself from Korp in several ways. Most notably, in Strix, a search result corresponds to a document rather than an individual occurrence. Furthermore, Strix includes additional features such as support for metadata filtering, text similarity analysis, and a reading mode with annotation highlighting.

5 Utilizing the Corpus Infrastructure

This section provides a detailed account of the data annotation process and the subsequent steps taken to ensure its accessibility through Korp and Strix. As previously mentioned, Språkbanken’s corpus infrastructure utilizes a pipeline architecture, which is known as Sparv. This pipeline encompasses a variety of both external and internal annotation tools designed for a wide range of word, structural, and text-level annotations. A comprehensive list of available annotations and the corresponding annotation tools are listed in the Sparv user manual¹¹.

⁸https://spraakbanken.gu.se/korp/#?stats_reduce=word

⁹<https://spraakbanken.gu.se/>

¹⁰<https://spraakbanken.gu.se/strix>

¹¹<https://spraakbanken.gu.se/sparv/#/user-manual/available-analyses>

Category	# Documents	# Sentences	# Words
Berättelser och redogörelser (narratives and accounts)	803	3,041,471	61,348,401
Betänkanden, memoria, och utlåtanden (reports, memorandums and opinions)	49,919	7,767,619	195,467,124
Motioner (motions)	77,555	3,324,913	73,189,180
Propositioner och skrivelser (propositions and letters)	19,761	13,516,498	319,201,218
Protokoll (protocols)	10,653	13,092,772	327,554,657
Register (register)	1,262	1,847,991	23,323,395
Reglementen (regulations)	133	129,154	2,628,009
Riksdagens författningssamling (the constitution of the Riksdag)	52	3,880	83,964
Riksdagsskrivelser (letters of the Riksdag)	30,383	1,359,009	29,775,566
Statens offentliga utredningar (government official investigations)	701	2,369,348	59,266,835
Total	191,222	46,452,655	904,138,844

Table 1: Statistics on the number of documents, sentences, and words category-wise and in total in the corpus collection.

Word-Level Annotations		Text-Level Annotations	
Annotation	sense: marknad	Annotation	Blingbring:innanhav
	compound word forms:[empty]		Swedish FrameNet: Amounting_to
	compounds:[empty]		readability index:50.09
	universal features:		
	Case:Nom		
	Definite:Ind		
	Gender:Com		
	Number:Sing		
	dependency relation: Complement of preposition		word variation index:66.63
	sentiment:neutral		nominal quote:2.21
	Swedish FrameNet: [empty]		file name: kombet_1911____6_5-D8E06b5b1.txt
	baseform:marknad		category:utredningar-kombet-sou
	lemgram:marknad (noun)		date:1911
	msd (Morpho-Syntactic Analysis):		
	NN.UTR.SIN.IND.NOM		
	part-of-speech:noun		

Table 2: List of word and text level annotations.

Leveraging Sparv, we have substantially enhanced our dataset with a diverse array of annotations. Table 2 provides an overview of various word and text-level attributes and their respective values illustrated through an example search term, 'marknad,' within the sentence: *Den mesta strömmingen föres saltad till marknad i Linköping.* (Most of the herring is brought salted to market in Linköping)

Some of these attributes are self-explanatory, while others need a brief explanation. The 'universal features' attribute reveals essential linguistic details such as case, definiteness, gender, and number associated with the selected word. The 'dependency relation' attribute provides the dependency relationship of the selected word with the head in the sentence. The 'sentiment' indicates the emotional tone. More details can be found in the Sparv user manual as mentioned above.

Text-level annotations encompass critical metadata, including the corpus category, file name, date, and some other text-level attributes.

Regarding corpus search functionality, Korp of-

fers three distinct types of searches: 'simple,' 'extended,' and 'advanced.' The 'simple' search is the most straightforward free-text search, allowing users to query the corpus for any given string. Figure 1 illustrates the search hits in the Keyword In Context (KWIC) view when searched for the term 'marknad' (ENG *market*). The left-hand pane displays sentences retrieved from all documents within selected corpora containing the search term, while the right-hand pane presents both text-level and word-level attributes (for the selected string highlighted with a black background) as explained above.

The search can be refined or broadened by utilizing the 'Extended' search tab, which permits filtering based on various word and text-level attributes, and combining various attributes with logical AND and OR operators. For instance, one can filter the sentences where the term 'marknad' occurred as a noun only (or any other part-of-speech tag and other attributes).

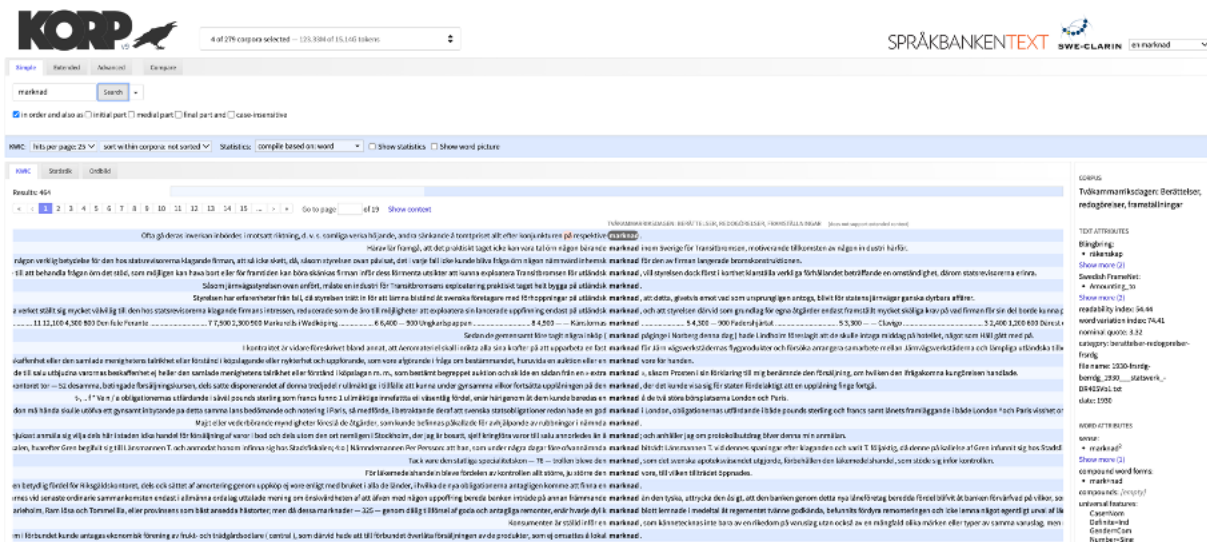


Figure 1: Screenshot of Korp frontend 'Basic' search

Additionally, for more specialized requirements, the frontend offers an 'Advanced' search option, enabling users to design search queries using the CQP query language (Christ, 1994). Beyond these functionalities, the frontend comes with various other compelling features, including the ability to display context (for or after the searched term), which can be invaluable during corpus exploration.

In addition to presenting the search hits in the KWIC view, the statistics on total number of occurrences for each matched word in the selected corpora overall and in each of the sub-corpora separately are also available under the 'statistics' tab (see Figure 1). Another useful view is the 'word-picture' view which shows other words associated with the searched word based on certain dependency relations such as subject, object, preposition, pre and post modifiers etc. Figure 3 shows word picture of the word 'marknad' within the 'Riksdag' corpora. Such a view can come handy while analyzing a search word within a corpus by looking at associated words or by comparing it to another corpora (or sub-corpora). For example, in Korp we also have another corpora of the recent Swedish parliament proceedings (Riksdagen-open-data). The word 'marknad' can be compared in these two collections through the word-picture view. Figure 4 present this comparison, and as can be noted, the separate list of associated words in two collections can reveal certain comparative aspects of the corpora. For example, while the bicameral data speaks of the market as Joint, Nordic, domestic, European,

foreign, open, large and free, the modern parliamentary data speaks of it with a different focus on free, joint, global, open, black, digital, international and lucrative.

Due to space constraints, we cannot delve into all of Korp's features here, so we encourage readers to visit the <https://spraakbanken.gu.se/korp/> to explore the corpus and experiment with the many search options available.

As can be seen in the given screenshots, in Korp each search hit is restricted to 'a sentence' (or a few sentences if the context visualization is turned on). An alternative is to return the documents containing the searched terms as search hits (as opposed to sentences), and then provide an option to view the full document in reading mode. This is exactly, what the Strix tool and interface is designed for. If we search for the term 'marknad' through the Strix interface, a list of documents from the collection containing the search term will be displayed as shown in Figure 5.

This list can be filtered further based on various text-level attributes (e.g. document type, document title, year etc.) using the given metadata filtering options in the left-hand side pane.

Clicking on any document will open the full document in text mode as shown in Figure 6.

Various text and word-level attributes of the selected text are displayed in the right-hand side, while the document itself is displayed on the left

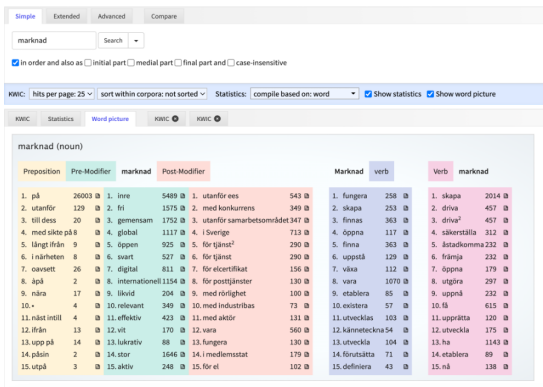


Figure 2: 'marknad' in the modern corpus

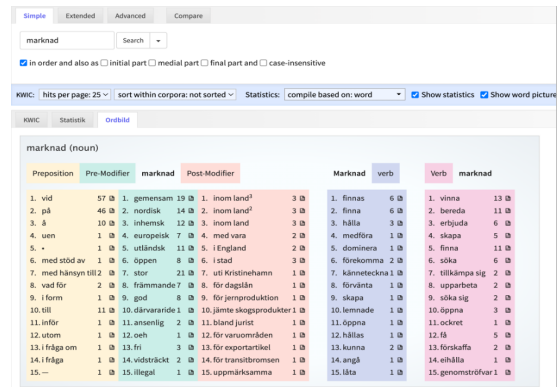


Figure 3: 'marknad' in the bicameral corpus

Figure 4: The word-pictures view in Korp

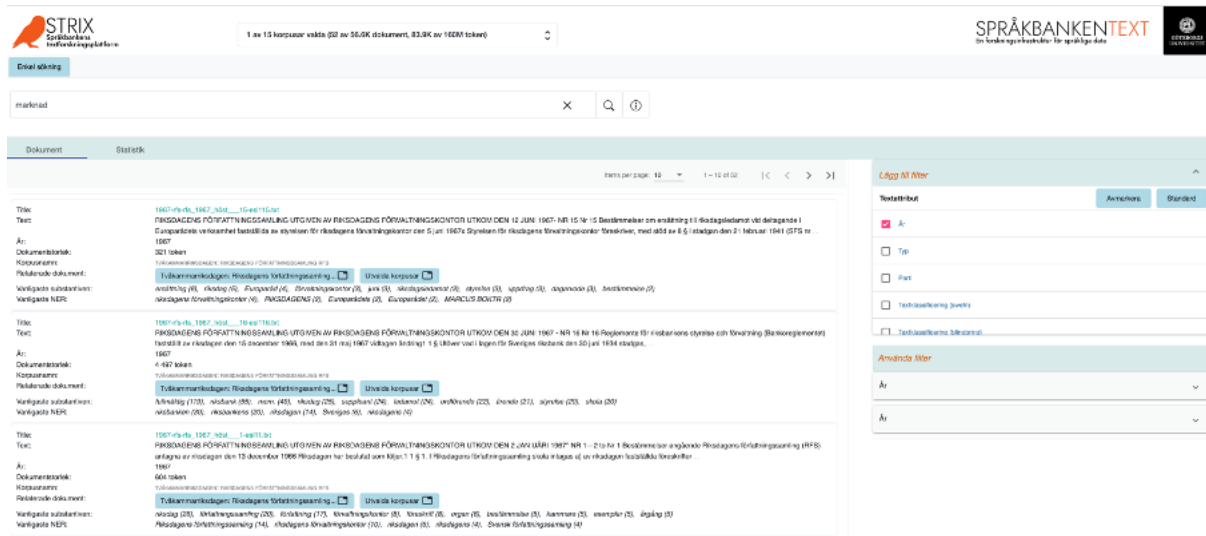


Figure 5: Screenshot of Strix

and side pane. Also note that the selected document can be further searched using the 'Search the current document' search box on top. Again, due to the space limitations, it is not possible to explain all searching and exploring options provided by Strix, and we refer the reader to Språkbanken for further details.

5.1 Resource URL's

The following url can be used to reach the Korp:
<https://spraakbanken.gu.se/korp/>

After switching to the 'English' mode, a particular corpus (or sub-corpus) can be selected using the drop-down list of available corpora before making the search as shown in Figure 7. Note, the 'bicameral Riksdag' corpus is placed under the 'Governmental texts' category.

Use the following url to access the Strix and then select the 'bicameral Riksdag' corpus from the drop-down list:

<https://spraakbanken.gu.se/strix/>

Once opened, add filters (if any) in the right-hand side pane, and then make the search as explained above.

6 The Market Language

The research project "The Market Language" has particularly benefited from the inclusion of material from the bicameral parliament in the Swedish Language Bank collection of corpora. The project's need for corpus data spanning different time periods and sources has also played a role in the decision to process this material.

"The Market Language: Studies in the Discourse about Markets from Medieval to Modern Times"

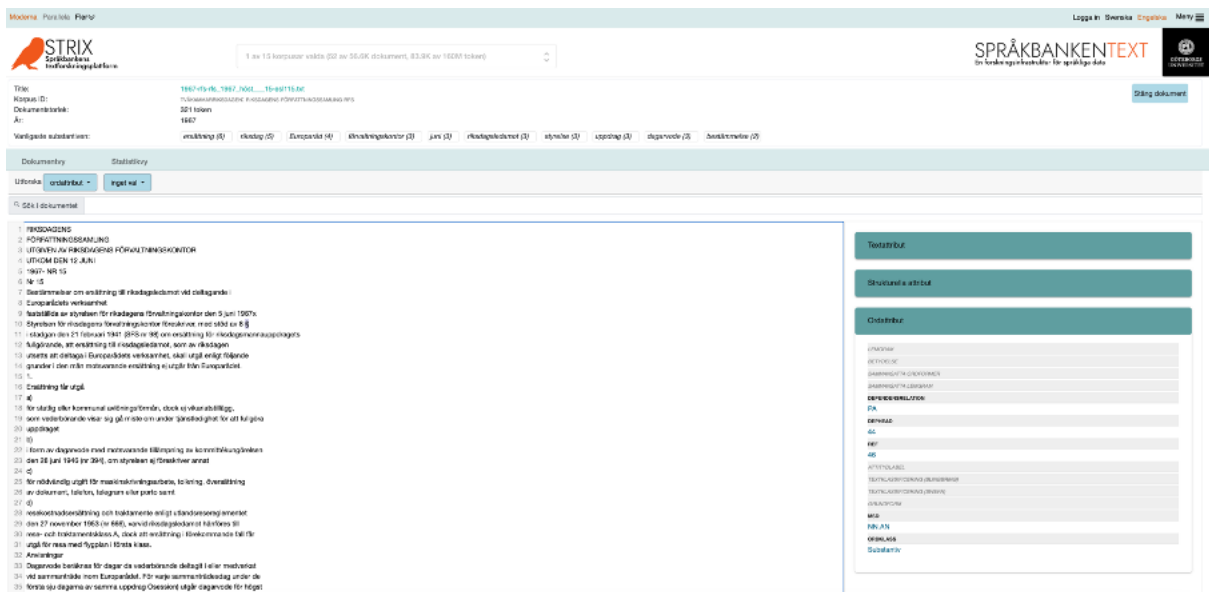


Figure 6: Strix Document View

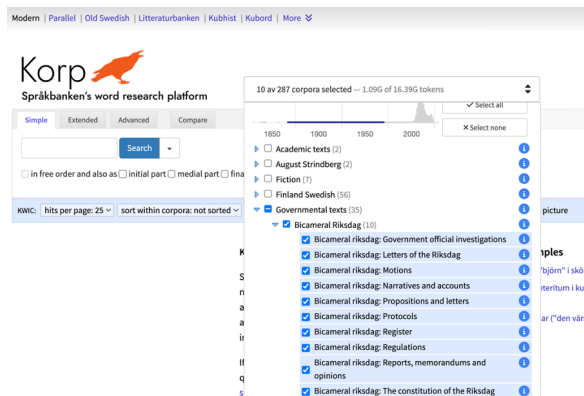


Figure 7: List of available corpora through Korp

is a four-year multidisciplinary project involving scholars from history of ideas, economic history, discourse-oriented linguistics, and computational linguistics. A central goal of the project is to deepen our understanding of the historical development that has led to the market becoming such a dominant concept today. This is achieved by analyzing how the word "market" has been used in the Swedish context over an extended period. A key objective of the project is to describe how expressions associated with the concept of "market" have been employed in Swedish language from the Middle Ages to the present. From the outset, the project aimed to combine qualitative methods from conceptual history and linguistic discourse analysis with primarily quantitative corpus linguistic analysis.

Making the complete records of the bicameral

parliament from 1867-1970 available in the Language Bank's frontend has given the project's collaborators numerous new opportunities to conduct searches without needing specific computational linguistic expertise. A notable advantage is that the material can now be used and processed for searches based on linguistic form or language functions in text. While searchable text files were available previously, their format severely limited the ability to observe how specific lexical items, such as words or phrases denoting the market phenomenon, were expressed over time. This is particularly important because the material reflects political discourse at a time when both domestic and international economies were becoming increasingly interconnected, and the meaning of "market" was expanding.

One focus of the project has been to examine how the original, concrete meaning of "market" as a place or location in time has increasingly been replaced or complemented by abstract meanings of "market." By being able to search the bicameral parliament material in the Korp interface, all researchers in the project team, with different levels of computer linguistic training, have been able to confirm working hypotheses and develop new searches about how the market as an abstract concept has been expressed. For instance, a simple comparison of the occurrences of "market" in its indefinite form (marknad) with its definite form (marknaden) reveals a tendency for the definite form to be used more frequently over time. This

suggests that "market" began to be referred to as an entity with increasing agency during this period, which is interesting when compared with its usage in modern corpus material, where the definite form is used to a higher degree. This indicative result can be followed up by finding cases where "market" is used in the subject role in a sentence as resulting from an intra-team discussion and collaboration based on the initial search results. Figure 9 shows an increased use of market as a subject, both in the indefinite and definite forms with a sharper increase of the definite form *marknaden*.

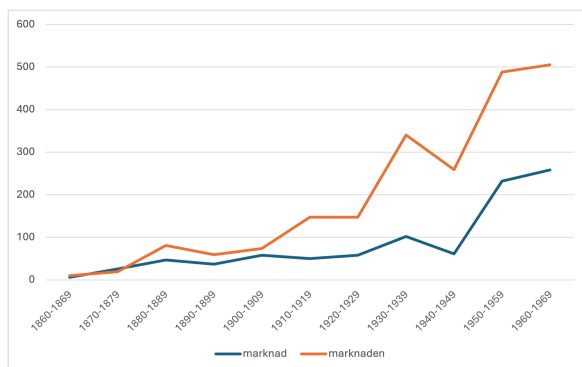


Figure 8: distribution of indefinite and definite forms of "market" over time 1860–1970 in absolute frequency

In this way, the parliamentary material from 1867-1970 serves as a basis for observing the evolution of currently established meanings.

Another clear example of the benefits for the project of including the bicameral parliament material in the corpus infrastructure is the ability to map and analyze compounds containing "market." Forming multi-element compound words is a typical feature of word formation in Swedish and other Germanic languages with English as a notable exception. Through annotation work and the various functions of the Korp interface, it is easy to list all compounds with "marknad" as an element, both when it serves as a prefix and as a suffix. The project team has conducted many such mappings, and these searches reveal a great variation in the meaning and discursive semantics of these compounds. This type of search can now easily be conducted by all team members in the Korp interface, which facilitates initial interpretation and discussion of results. One more specific result is that new compounds with "market" appear to have been continuously formed over the period covered by the material, often in relation to the

political changes discussed in parliament at different times. Compounds with concrete connotations (like "cattle market" or "butter market") are common in the mid 1800s but are gradually replaced by increasingly abstract constructions that indicate the Swedish economy's connection to other economies (such as "world market" or reference to a "domestic market" in contrast to "foreign markets"). Furthermore, the concept of "market" also extends into other political areas (such as "labor market" or "housing market") during the 1900s. Figure 9 shows the increase of the compound word "arbetsmarknad" (labor market) in relative frequency in the parliamentary texts over time. The compound is most frequent towards the end of the bicameral parliament in Sweden in the 1950s and 1960s.

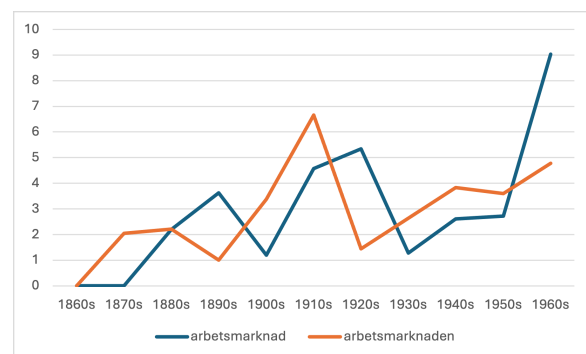


Figure 9: distribution of relative frequency (per million) occurrences of forms for "labor market" over time

7 Conclusions

We have annotated the Swedish bicameral parliament data with text-level as well as token-level attributes to make searching, filtering, and exploration much easier and more useful for broader groups of potential users with different level of technological knowledge. The Market Language project has incentivized the inclusion of the dataset into the infrastructure. The research team is now able to work on the project's original research questions and can also address new questions that arise from the results with the help of the infrastructure interface. This makes it easier for team members without computational linguistics training and also improves the quality of the output data. Further, the inclusion of bicameral parliament data in the Språkbanken environment helps complete the picture of making public texts in Swedish available,

both for the research community and for other public purposes.

We believe and hope that this collection of texts will be a valuable resource for deeper analysis of the Swedish political discourse during the period when the bicameral parliament was in function. Together with existing corpora of other parliamentary data and government texts, the bicameral parliament corpus will also enable more informed cross-corpora searches. This benefits scholars in historical studies as well as researchers from many other disciplines, not least in studies of language. The data of the bicameral parliament has existed in other formats for a long time, but the inclusion of this material into Språkbanken makes it accessible and also adds a valuable new component to this infrastructure.

Acknowledgements

This work has in part been funded by the research program *Change is Key!* supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. *Korp — the corpus infrastructure of språkbanken*. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 474–478, Istanbul, Turkey. European Language Resources Association (ELRA).
- Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. *ArXiv*, abs/cmp-lg/9408005.
- Peter De Bolla. 2023. *Explorations in the Digital History of Ideas: New Methods and Computational Approaches*. Cambridge University Press.
- John Patrick Leary. 2019. *Keywords: The new language of capitalism*. Haymarket Books.
- Claes Ohlsson, Victor Wählstrand Skärström, and Henrik Björck. 2022. The market as a concept in swedish parliamentary records from 1867 to 1970: A mixed methods study. In *Digital Parliamentary Data in Action (DiPaDA 2022) workshop, Uppsala, Sweden, March 15, 2022*, pages 22–34. CEUR-WS. org.

Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams

Roberts Dargis, Guntis Bārzdīns, Inguna Skadiņa, Normunds Grūzītis, Baiba Saulīte

Institute of Mathematics and Computer Science, University of Latvia

Raina bulv. 29, Riga, LV-1459, Latvia

{roberts.dargis, guntis.barzdins, inguna.skadina, normunds.gruzitis, baiba.saulite}@lumii.lv

Abstract

The latest large language models (LLM) have significantly advanced natural language processing (NLP) capabilities across various tasks. However, their performance in low-resource languages, such as Latvian with 1.5 million native speakers, remains substantially underexplored due to both limited training data and the absence of comprehensive evaluation benchmarks. This study addresses this gap by conducting a systematic assessment of prominent open-source LLMs on natural language understanding (NLU) and natural language generation (NLG) tasks in Latvian. We utilize standardized high school centralized graduation exams as a benchmark dataset, offering relatable and diverse evaluation scenarios that encompass multiple-choice questions and complex text analysis tasks.

Our experimental setup involves testing models from the leading LLM families, including Llama, Qwen, Gemma, and Mistral, with OpenAI's GPT-4 serving as a performance reference. The results reveal that certain open-source models demonstrate competitive performance in NLU tasks, narrowing the gap with GPT-4. However, all models exhibit notable deficiencies in NLG tasks, specifically in generating coherent and contextually appropriate text analyses, highlighting persistent challenges in NLG for low-resource languages.

These findings contribute to efforts to develop robust multilingual benchmarks and to improve LLM performance in diverse linguistic contexts.

1 Introduction

The dream that artificial intelligence (AI) can perform many tasks in a similar manner to humans became closer with the release of ChatGPT by OpenAI in November 2022. Today, several large language models (LLM) have been made available by global companies and are widely used by society

and industry for various text generation tasks, such as question answering, text summarization, translation, etc. However, LLMs have shown considerably less reliable results for low-resource languages (Lai et al., 2023; Ahuja et al., 2024). The reason for this is the fact that most of the language data used for training LLMs is in English and few other widely spoken languages, while low-resource languages are represented by very small portions of data.

Benchmarking is a crucial step in evaluating LLM performance and capabilities across various tasks. It involves setting standardized tests or tasks to measure the LLMs' performance. A lack of benchmarks that enable comprehensive multilingual evaluation is one of the reasons why research on LLMs and machine learning models for NLP is still mostly focused on English and some other widely spoken languages.

The aim of this paper is to conduct an initial evaluation of open-weights LLM capabilities in Latvian, both in natural language understanding (NLU) and in natural language generation (NLG). The evaluation was performed using high school centralized graduation exams, overseen by the National Centre for Education. High school exams serve as an excellent benchmark dataset because they offer a relatable point of reference, allowing for comparison not only between different models but also between the performance of LLMs and the expected achievements of high school graduates.

2 Related Work

Recent advances on LLMs have led to impressive gains on NLU benchmarks, starting from GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) with 10 tasks related to different NLU problems, followed by MMLU (Hendrycks et al., 2021) which covers nearly 60 subjects (including STEM: science, technology, engineering and mathematics), and Bigbench (Srivastava et al., 2023) with more

than 200 tasks, as well as many other benchmarks.

Again, many well-known benchmarks are available only in English and other widely spoken languages. Google researchers addressed the need for a highly multilingual benchmark when the first transformer-based LLMs appeared by introducing the Cross-lingual Transfer Evaluation of Multilingual Encoders (XTREME) benchmark (Hu et al., 2020) which is used to evaluate cross-lingual generalization capabilities of multilingual representations. Although XTREME covers 40 typologically diverse languages, spanning 12 language families, Baltic languages are not included in this benchmark. Similarly, the dataset for the evaluation of multilingual LLMs developed by Okapi (Lai et al., 2023), in which the English part was translated with the help of ChatGPT, covers 26 languages except the Baltic languages (the “smallest” language is Danish with 6 million speakers, followed by Slovak with 7 million speakers).

The development of test sets for specific languages involves significant human resources. Therefore a widely used strategy is to apply machine translation, with or without manual post-editing. Recently, this approach was chosen to translate the MMLU and COPA (Gordon et al., 2012) datasets into Latvian.¹ Evaluation of OpenAI ChatGPT 3.5 Turbo and Google Gemini 1.0 Pro on the machine translated MMLU dataset shows that performance of these LLMs for Latvian is worse when compared to English (Bakanovs, 2024). It should be noted, that this dataset is not manually post-edited, and machine translation most likely has introduced some errors which can result in generating wrong answers. Bakanovs (2024) experiment on a small subset of the social science domain shows that post-editing improves results by 3 percentage points for ChatGPT-3.5 Turbo and by 9 percentage points for Gemini 1.0 Pro.

Finally, GPT-4 has been evaluated by OpenAI on several benchmarks (OpenAI et al., 2024), such as MMLU, HellaSwag, AI2 Reasoning Challenge, WinoGrande, HumanEval, and DROP. When comparing GPT-4’s 3-shot accuracy on MMLU across different languages, English reaches 85.5% (only 70.1% for GPT 3.5), while Latvian – 80.9%. With respect to educational tests and exams, OpenAI has reported that “GPT-4 exhibits human-level performance on the majority of professional and aca-

demic exams” (OpenAI et al., 2024).

3 Test Setup

All tests were run using the Ollama toolkit on a computer with 8x interconnected Nvidia A100 80GB GPUs.

The most popular open-source LLM families were chosen to be tested: Llama, Qwen, Gemma, and Mistral. A non-quantized *instruct-fp16* version was chosen for each model, except for Llama3.1 405B because the model was too large, therefore its 5-bit K-quantized version was used instead.

The emphasis in this article is on open-weights models. OpenAI’s GPT-4o model is added just for a reference as the most popular closed-source commercial model. In the GPT-4 technical report (OpenAI et al., 2024), Latvian is classified as a low-resource language. Although it could be argued that Latvian is not as low-resource as many other languages, especially w.r.t. to the number of native speakers, it is considered as low-resource also by European researchers (Ali and Pyysalo, 2024).

4 Centralized High School Exams

In Latvia, centralized exams are a crucial component of the educational system, designed to standardize knowledge assessment across the country and to ensure that high school graduates meet national academic standards. These exams are taken at the end of the 11th or 12th grade and are required to obtain a high school diploma.

Students in Latvia must take a certain number of centralized exams, though they have some flexibility in choosing which subjects to be examined in, depending on their future academic and career aspirations. It is not expected for a student to be able to pass the exams in all subjects. The mandatory exams include Latvian language and literature, mathematics, and a foreign language of choice (usually English, but alternatives such as German, Russian, or French are available). Beyond these core subjects, students can opt to take additional exams in subjects like biology, chemistry, physics, history, geography, or informatics.

Higher education institutions in Latvia typically use these scores as part of their admission criteria, often alongside other considerations such as entrance exams or interviews. This makes the performance on centralized exams a significant factor in a student’s educational trajectory.

These exams are designed and administered by

¹The Latvian versions of these datasets are available at <https://github.com/LUMII-AILab/VTI-Data>

Model	Val.	Con.	Corr.
gpt-4o	1.00	0.88	0.82
gpt-4o-mini	1.00	0.86	0.78
llama3.1 : 405b	0.99	0.75	0.72
qwen2 : 72b	1.00	0.89	0.72
llama3 : 70b	1.00	0.88	0.71
gemma2 : 9b	1.00	0.89	0.68
gemma2 : 27b	0.97	0.90	0.67
llama3.1 : 70b	1.00	0.72	0.64
mistral-large : 123b	0.99	0.71	0.63
gemma2 : 2b	0.97	0.71	0.40
qwen2 : 7b	0.97	0.64	0.40
llama3 : 8b	0.92	0.43	0.31
llama3.1 : 8b	0.93	0.32	0.26
mistral-nemo : 12b	0.10	0.00	0.00

Table 1: LLM performance on MCQ tests in Latvian. Val. – validity; Con. – consistency; Corr. – correctness.

the National Centre for Education.² The exams are intended to assess not just rote memorization, but also critical thinking, problem-solving abilities, and application of knowledge. The structure of the exam and the types of tasks vary from year to year.

Exams of 2023 were chosen for the initial version of this benchmark, since a lot of the exercises contained multiple-choice questions (MCQ). In addition, models were also tested on text analysis task from the Latvian language and literature exam.

5 Multiple-choice Questions

A set of 72 Latvian MCQs was created, covering physics, geography, chemistry, biology, Latvian language and literature exams. Questions containing pictures and complex formulas were omitted.

The models were tested with the zero-shot learning approach. The prompt started with a question, followed by answer options and concluded with the instruction: “Atbilde formātā ‘Atbilde ir X’, kur X ir pareizās atbildes burts.” (“Answer in the form ‘Answer is X’ where X is the letter of the correct answer.”). The results are shown in Table 1.

The first criterion evaluated was validity – how many of the generated answers matched the expected format. Many models achieved 100% validity, indicating that instructions were understood and the zero-shot approach works well for this kind of task. For an answer to be valid it must contain the phrase “Atbilde ir” followed by a letter A–Z.

²Past exams are available at <https://www.visc.gov.lv/lv/valsts-parbaudes-darbi>

There can be any number of whitespaces and asterisks (used by some models to indicate bold text in the markdown syntax) between the phrase and letter. The upper/lower case of letters is ignored. The answer may also contain extra text (usually, an explanation) before or after the phrase.

The second criterion evaluates consistency. Each prompt was sent to each model 10 times with a different seed value each time to reduce the chance of a lucky guess. The model must choose the same option for the same question every time. This criterion was evaluated on a per-question basis. To count an answer to a question as consistent, all answers to the same question must be valid, and the chosen option must be the same in every attempt.

The final measure binds it all together. For a question to be counted as correctly answered, all responses must be valid, consistent, and correct. Such a strict requirement was used to measure the true expected correctness rate. The questions in biology and geography had higher correctness scores overall. The chemistry scores were lower because some of the questions contained chemical reaction equations, and some physics questions required not only reasoning, but also calculations.

The non-quantized *fp16* models had very similar correctness compared to 5-bit K-quantized models. For such tasks, the quantized models would be more appropriate due to their significantly smaller memory and compute footprint.

6 Text Analysis and Writing Skills

One of the tasks in the Latvian language and literature exam in 2023 was to read two texts (each about 600 words) and write a text analysis (500–600 words) comparing both texts, following the principles of text composition and including the specified content components:

- Topic, relevance, and issues.
- Cultural facts, signs, or symbols in the interpretation of the cultural-historical context.
- Connection with other cultural facts beyond the provided texts.
- Text composition, genre characteristics.
- Language tools typical to the author’s style in the analyzed texts.

The same task was given to the largest model from each of the LLM families. The result was

Model	Understanding (4-16)	Argumentation (3-12)	Language (0-16)	Creativity (4-16)	Total (11-60)
gpt-4o : 2024-08-06	12	9	14	11	46
gemma2 : 27b	12	9	11	10	42
llama3.1 : 405b	9	7	13	9	38
mistral-large : 123b	11	9	7	7	34
qwen2 : 72b	4	3	5	4	16

Table 2: Human expert evaluation of LLM text analysis and writing skills.

evaluated by an expert in linguistics using the same guidelines and criteria as students were evaluated on the exam. The results are evaluated according to 15 criteria divided into four categories: knowledge and understanding, argumentation, language quality, and creativity. For each criterion, students can get 1 to 4 points, except for language quality for which 0 points can be assigned as well. The overall score can range from 11 to 60 points. In our evaluation, the same scoring method was used to strictly comply with the official guidelines.

The results of the evaluation of the text analysis and writing skills are shown in Table 2. GPT-4o and Mistral are the only models that generated text within the requested length (500–600 words). Gemma2 generated almost 500 words, while Qwen2 and Llama3.1 generated about 250 words. The language quality of Llama3.1 and Gemma2 was very similar. According to the guidelines, language quality is based on absolute number of errors, thus, comparing two texts of similar relative quality, the longer one typically will have more errors and therefore a lower score.

Similarly, knowledge and understanding was based on the number of facts mentioned in the text, therefore shorter analysis had a disadvantage in this category.

The text generated by Qwen2 was very difficult to understand with many illogical sentences, which led to a low score in other categories. The text generated by Mistral had many agreement errors, such as subject-verb, noun-adjective, tense, gender, and singular/plural disagreement.

Demonstrating author’s individuality was one of the conditions to get top scores in originality (part of creativity), which was lacking in all of the analysis. There was also a lack of comparison to nowadays, which was a condition to get top scores in the knowledge category.

All of the models analyzed the two texts mostly separately, using the specified content component

subsections. A cohesive, fluent analysis with introduction, discussion, and conclusion was expected instead. In this task, zero-shot learning did not work well. For such tasks, examples or more detailed instructions provided in the prompt would probably lead to better results.

Experiments were also conducted using LLMs as evaluators. Each model was asked to assess the text analysis generated by every other model ten times. The results were not promising. The scores varied a lot between the runs, and the average scores by any model did not correlate with human evaluation. The worst category is language quality assessment. Most of the errors found by the models were not actually errors, and many actual errors were missed.

7 Conclusion and Further Work

The experiments validated the use of Latvian high school centralized graduation exams as a source for natural language understanding (NLU) tasks. This gives us motivation to continue the work on expanding the size of the data set. The performance gap for Latvian between the best open-weights LLMs and GPT-4o is minor. The biggest surprise was the performance of the relatively small Gemma2 27B model. The quantized version is small enough to be run on a consumer grade GPU, making it perfect for large-scale NLU tasks, such as classifying or tagging documents even without fine-tuning. This opens up huge possibilities for NLP in Latvian, especially in digital humanities.

The performance in natural language understanding did not correlate with the performance in natural language generation (NLG) in text analysis tasks, showing the importance of evaluating both tasks separately. Despite NLG shortcomings discussed, in human evaluation the best open-source LLMs achieved score above 66% (40 out of max 60 points) compared to average 56% score reported for the actual human graduation exam. Unfortu-

nately, NLG tasks are hard to evaluate, since human evaluation requires a lot of resources. Even the best models showed no correlation between their assessment of other LLM’s on the NLG tasks, and the human evaluation. This makes the current generation of LLMs not well suited for NLP tasks like error detection and correction, text normalization and data denoising in Latvian.

The high out-of-vocabulary (OOV) word density score, measured against the large Latvian Thesaurus database (Grasmanis et al., 2023) was a good indicator of poor language quality, but a low number of OOV words is not an indicator of high NLG score, because most errors were grammatical errors. Finding a good automatic NLG evaluation methodology is still an open research question.

The dataset created and used in this evaluation is available as open data via a GitHub repository.³

Overall, the open-weights models show promising performance on Latvian, suggesting that fine-tuning such models for low-resource languages might achieve competitive results with much lower costs compared to training language-specific LLMs from scratch.

Acknowledgments

This work was supported by the EU Recovery and Resilience Facility project Language Technology Initiative (2.3.1.1.i.0/1/22/I/CFLA/002) in synergy with the Latvian Council of Science grant Common Writing Errors in Latvian (Izp-2023/1-0481).

References

- Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. [Megaverse: Benchmarking large language models across languages, modalities, models and tasks](#). *Preprint*, arXiv:2311.07463.
- Wazir Ali and Sampo Pyysalo. 2024. [A Survey of Large Language Models for European Languages](#). *Preprint*, arXiv:2408.15040.
- Bruno Bakanovs. 2024. Large Language Model Evaluation and Improvements for the Latvian Language. Master’s thesis, University of Latvia.
- Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. [SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning](#). In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 394–398. Association for Computational Linguistics.
- Mikus Grasmanis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma, Laine Strankale, Arturs Znotins, and Normunds Gruzitis. 2023. [Tēzaurus.lv – The Experience of Building a Multifunctional Lexical Resource](#). In *Electronic lexicography in the 21st century (eLex): Invisible Lexicography*, pages 400–418.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring Massive Multitask Language Understanding](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization](#). *CoRR*, abs/2003.11080.
- Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. [Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, et al. 2024. [GPT-4 Technical Report](#). *Preprint*, arXiv:2303.08774.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adria Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, et al. 2023. [Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

³<https://github.com/LUMII-AILab/VTI-Data>

Computational Methods for the Analysis of Complementizer Variability in Language and Literature: The Case of Hebrew *še-* and *ki*

Avi Shmidman

Bar-Ilan University

Dicta: The Israel Center for Text Analysis

avi.shmidman@biu.ac.il

Aynat Rubinstein

The Hebrew University

of Jerusalem

aynat.rubinstein@mail.huji.ac.il

Abstract

We demonstrate a computational method for analyzing complementizer variability within language and literature, focusing on Hebrew as a test case. The primary complementizers in Hebrew are *še-* and *ki*. We first run a large-scale corpus analysis to determine the relative preference for one or the other of these complementizers given the preceding verb. On top of this foundation, we leverage clustering methods to measure the degree of interchangeability between the complementizers for each verb. The resulting tables, which provide this information for all common complement-taking verbs in Hebrew, are a first-of-its-kind lexical resource which we provide to the NLP community. Upon this foundation, we demonstrate a computational method to analyze literary works for unusual and unexpected complementizer usages deserving of literary analysis.

1 Introduction

Natural languages offer speakers a variety of means for expressing their sentiment and attitude toward events, be they actualized or unactualized. While the literature has traditionally focused on lexical items that convey sentiment and attitude (verbs, adjectives, nouns; underlined in (1)-(2)), it is well known that functional morphemes such as subordinating particles (*complementizers*; bold-faced in the examples) and mood inflection are also implicated in the expression of such meanings in certain languages (Mauri and Sansò, 2016).

- (1) We {are proud / believe} **that** our athletes did their very best.
- (2) ha-tiqva **še-/ki** taxzeru mexazeqet.
the-hope COMP you.will.return strengthens
'The hope that you all will return is emboldening.' (Hebrew)

Recent work on the interaction between attitude predicates and the grammatical forms they “select” in the embedded clause has pointed to subtle

semantic effects of choosing one complementizer over another (in Greek; Giannakidou and Mari 2021), or one inflected form over another (in Romance languages; Portner and Rubinstein 2020; Mari and Portner 2021). In contemporary Hebrew, the language we focus on in this paper, the variation between *še-* and *ki* as complementizers has not been recognized as relating to the grammar of embedding, and has often been attributed to register: *ki* is viewed as being more formal (see Nir 2013).¹

Understanding patterns of clausal complementation in a language and the allowed range of *variation* is crucial for both comprehension and production. The distinctions are subtle and may seem arbitrary. They are known to present substantial difficulty even for advanced second language (L2) learners (e.g., Bartning and Schlyter 2004 on French; Kanwit and Geeslin 2018 on Spanish).

This paper presents the first attempt we know of to explore aspects of complementizer distribution and use from a computational perspective, in Hebrew but also more generally. As we survey in Section 2, earlier computational studies of clausal embedding in attitude contexts have focused on English or on curated annotations. Corpus-informed studies have been limited to languages with mood inflection (e.g., French; Petkovic and Rabiet 2016), leaving complementizer variability unexplored. The contributions of the present paper are as follows:

- Enriched lexical representations of clause-taking verbs in Hebrew, with corpus-based statistics both regarding overall tendencies, as well as regarding the degree to which these tendencies are exaggerated or overridden in marked contexts.
- Demonstration of how this data can be lever-

¹*Ki* has an additional use as a subordinating conjunction of reason ('because'). We set it aside in what follows.

aged to reveal the characteristics of specific marked contexts which require selection of one or the other of the complementizers. The ability to identify these marked contexts is a key component for L2 instruction.

- Presentation of a language-agnostic method to identify unusual usages of subtle linguistic elements in literary corpora.
- Application of this method to a corpus of modern Hebrew literature, identifying unusual specimens that invite further literary analysis.

2 Related work

Large-scale datasets allowing for the investigation of clausal embedding have been developed within the MegaAttitude Project.² In particular, the MegaAcceptability dataset (White and Rawlins, 2020) provides acceptability judgments on the distribution of 1,000 attitude verbs in 50 syntactic frames in English. Additional datasets explore inferences patterns associated with attitude verbs and their interaction with elements such as negation and tense (Moon and White 2020; Kane et al. 2021).

Özyıldız et al. (2023) provide a database of theoretically informed syntactic and semantic properties of a set of 50 attitude verbs in 15 languages. The rich linguistic profile of each verb (including its complement types, factuality inferences, interaction with negation, focus sensitivity, gradability and more) is summarized in a table based on experts' judgments in response to a questionnaire. Hebrew is included in this database, but information about complementizer variability is lacking from its description. Moreover, the database provides information about a small set of verbs and is based on translation from English, not on language-internal distributions.

Computational resources for languages that have observable mood inflection in embedded environments include mood as a target of morphological annotation (e.g., the Romance Verbal Inflection Dataset by Beniamine et al. 2020; Romance languages in Özyıldız et al. 2023). Petkovic and Rabiet (2016) provide a corpus-based description of mood variation in embedded clauses in French. However, languages in which mood is marked on a subordinating particle have

²<http://megaattitude.io/>

not yet received attention from a computational perspective.

In the NLP literature on Hebrew, Fadida et al. (2014) provide a corpus-informed dictionary of about 3,000 verbs along with the number and type of complements they tend to occur with, including clauses. The two complementizers *še-* and *ki* were treated as interchangeable in this earlier work. Our work extends this lexicon with detailed information about the complementizer variability characteristic of each verb.

The question of whether authors of Hebrew literature adhere to the same complementizer tendencies as found in other text types has not been previously explored. Nevertheless, we note that in one study of embedded clauses in Hebrew (Kuzar, 1993), examples of complementizer use from Hebrew literature are cited alongside those of newspapers, without any differentiation between the corpora.

3 Experimental Setup

3.1 Corpus-based Query for Complementizer Propensities

As noted above, different Hebrew verbs show propensities for different complementizer choices. However, these propensities have never before been investigated from a corpus perspective, due to the difficulty involved in running such a wide-scale query. Fortunately, such a query is now tractable. We use it to establish the overarching tendencies regarding complementizer use.

Hebrew corpus: We start with a corpus of 29 million modern Hebrew sentences, sourced from Hebrew news sources, Hebrew Wikipedia, Israeli Parliament Proceedings, and published Hebrew books (Table 1). We filter out sentences that are too short to be useful (under 6 words) or those that are overly long (over 60 words).³

Syntactic analysis: In order to identify cases of clausal complements, we obtain a syntactic parse for all of the sentences in our corpus. Currently, the two leading syntax parsers for modern Hebrew are the Levi-Tsarfaty parser (Levi and Tsarfaty, 2024) and DictaBERT-Parse (Shmidman et al., 2024). The former, though also achieving SOTA results, is not available to the public and is too

³In addition to this corpus, we also utilize a corpus of ten novels of the Hebrew novelist Amos Oz for a specific inquiry regarding Hebrew literature later in Section 6; see Appendix B for the list of books in that corpus.

Domain	Sentences (millions)
Hebrew Newspapers	22
Hebrew Wikipedia	5
Parliament Proceedings	1
Published books, Fiction	0.6
Published books, Non-fiction	0.4

Table 1: Sentences per domain in our Hebrew corpus.

slow to be tractable for a corpus as large as ours. We therefore used the latter. Parsing of the entire corpus required 73 hours on a single 4090 GPU.

In the parsed corpus, we examine all cases in which a verb has a "ccomp" dependency, indicating a clausal complement. We retrieve the complementizer by extracting the earliest token in the sentence within the scope of the clausal complement. The tabulation of the results of this query identifies the general tendencies of each verb in terms of complementizer choice.

3.2 Deviations from Complementizer Propensities

After calculating the overall complementizer propensities of each complement-taking verb, we wish to clarify the extent to which these propensities remain constant across usages of the verb, or whether there are specialized usages of the verb that exaggerate or override the general tendencies.

Especially interesting here is the question of whether certain usages or contexts entail exclusive use of one or the other of the complementizers. Received syntactic descriptions hold that either of the two complementizers can be used with virtually every clause-embedding attitude verb. However, in practice, there are cases where native Hebrew speakers will only find one of the two to be acceptable, while the other would sound unnatural.⁴ The method we present here allows us to pinpoint such cases.

Prima facia, in order to identify cases in which only one of the complementizers is used in practice, we might have considered simply running a BERT Masked Language Model (MLM) to see whether only one of the complementizers is predicted for a given context. However, in practice, given a masked token in place of the complementizer, BERT will almost always provide both of the complementizers among its top predictions,

⁴For examples of such, see Section 6 below.

because there simply aren't that many other options to fill the slot. That is, even if it would sound odd to a native speaker, if BERT's MLM head is pressed to choose a word to fill a complementizer function, and if the more usual complementizer has already been predicted, it will generally provide the other one, because, from a technical syntactic standpoint, both of them can theoretically function as a complementizer with any complement-taking verb. Instead, in order to gain a better sense of the extent of complementizer interchangeability, we examine contextualized embeddings for the complementizer positions, and we consider the extent to which the embeddings cluster into complementizer-specific sections, as follows:

Generating contextualized embeddings: For each complement clause identified in the previous step, we mask the complementizer, and we submit the sentence to BERT to generate a contextualized embedding for that masked token, independent of whether the complementizer was in fact *še-* or *ki*.⁵

2D Visualization: In order to visualize the interchangeability of the two complementizers across different contexts with the same governing verb, we reduce the 768-dimension space of the BERT embeddings using the t-SNE algorithm, and generate a two-dimensional plot of the embedding space for each complement-taking verb. We color the points based on the complementizer present in the corresponding sentence. As we demonstrate below, visual inspection of the relative distribution of the two colors across the plot allows us to easily and immediately identify areas of aberrations, representing specific contexts in which the tendency towards one complementizer or the other differs from the overall tendency within the corpus.

Clustering the embeddings: We add a clustering step to automatically isolate contexts with specialized complementizer tendencies. For each complement-taking verb, we collect the complementizer embeddings generated in the previous step (for practicality, we set a limit of 20,000 cases for any given verb; if the corpus contains more than this, then we randomly sample 20,000 cases

⁵In order to ensure that the embeddings are attuned to the nature of Hebrew prefixes (such as *še-*), we use a variation of the DictaBERT model. Leveraging the segmentation predictions of DictaBERT-Parse (Shmidman et al., 2024), we separate all prefixes in the DictaBERT training corpus into independent tokens, and then we run a new BERT pre-train based upon this prefix-separated corpus. The resulting BERT model is used to generate the embeddings for this step.

from across the corpus). We apply agglomerative clustering to these embeddings, with euclidean distance and average linking. We let the agglomerating continue until a majority of the samples have been clustered into the top three clusters. This ensures that the clustering process continues sufficiently long such that the majority usages of the word are clustered together in a few substantially-sized clusters, while still providing ample opportunity for specialized usages to occupy smaller individual clusters.

The key part of this clustering step is that neither the BERT embedding nor the clustering procedure has any information about the complementizer used in the sentence. This means that the algorithm cannot directly choose to cluster together sentences on the basis of the complementizer; rather, the clustering is based on the context alone. Thus, if a verb’s tendencies regarding complementizer usage are context-independent, then we expect the resulting clusters to each contain a mixture of *še-* and *ki* cases, reflecting the overall tendency of the verb towards one or the other. However, if we find clusters that are highly divergent from the overall tendency, this indicates that the types of contexts included in those clusters entail specialized complementizer tendencies.

In order to automatically evaluate the degree to which a given cluster diverges from the overall norm for the governing verb, we calculate the Jensen–Shannon Divergence (JSD) for each resulting cluster (discarding tiny outlier clusters of under 100 sentences); we consider a cluster to reflect a divergent complementizer tendency if it bears a JSD score higher than 0.04. On this foundation, we calculate, for each governing verb, the percentage of sentences that were clustered into divergent clusters. The result provides a measure of the extent to which the verb’s overall complementizer tendency holds true across the range of practical usages of the verb, in contrast with the extent to which the verb admits of specialized usages which affect its complementizer selection, and which an L2 learner would have to internalize in order to speak in a fully natural manner.

3.3 Pinpointing unusual usages

Finally, we wish to leverage the foregoing infrastructure to identify cases in which literary authors deviate from normative usage by choosing an unexpected complementizer, inviting literary anal-

ysis of the unusual choice. In order to do so, we run a set of modern Hebrew novels through the process above, isolating all cases of complement clauses, and generating contextualized embeddings for the complementizer in each case. We then use a K-nearest-neighbor classifier (with $k=3$) in order to classify each one of these cases according to the clusters for the corresponding verb that we produced in the previous step, based on the full large-scale Hebrew corpus. The cluster assignment provides us with a sense of the expectations for complementizer selection, given both the specific verb and the specific context of use. If the cluster assignment indicates a context in which one of the two complementizers is expected with a probability of over 95% (that is, a context in which the complementizers are effectively not interchangeable, but rather one is blocked in practical usage), and if the author nevertheless chose the *other* complementizer, then we flag the sentence as reflecting an unusual and unexpected complementizer choice.

4 Enriched Verbal Lexicon

Table 2 shows the overall proportion of the two complementizers for the ten most frequent clause-embedding verbs among the sentences analyzed. Tables 4-6 in Appendix A provide this information for the 100 most frequent clause-embedding verbs, along with additional statistical measures to be described below. A visualization of complementizer proportions is provided in Figure 1 for nine verbs. The wide range of complementizer variability observed across different Hebrew verbs has never before been quantified.

Verb	<i>ki</i>	<i>še-</i>
<i>amar</i> (‘said’)	55%	45%
<i>taʕan</i> (‘asserted’)	65%	35%
<i>xašav</i> (‘thought’)	4%	96%
<i>cuyan</i> (‘was mentioned’)	86%	14%
<i>hodiʕa</i> (‘informed’)	74%	26%
<i>qava</i> (‘decided’)	72%	28%
<i>yada</i> (‘knew’)	12%	88%
<i>heʕerix</i> (‘estimated’)	70%	30%
<i>siper</i> (‘told’)	51%	49%
<i>hevin</i> (‘understood’)	17%	83%

Table 2: Ten most frequent verbs with *ki/še-* clausal complements.

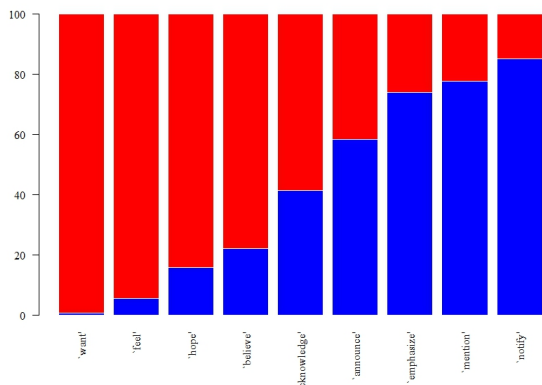


Figure 1: Proportion of *še-* (red) and *ki* (blue) complementizers for nine verbs. From left to right: *raca* (‘want’, 99% *še-*), *hirgiš* (‘feel’), *qiva* (‘hope’), *heʔemin* (‘believe’), *hoda* (‘acknowledge’), *hixriz* (‘announce’), *hidgiš* (‘emphasize’), *ciyen* (‘mention’), *masar* (‘notify’, 15% *še-*).

Previous literature noticed semantic trends in complementizer choice, but did not leave room for variation. For example, Zuckermann (2006) suggests the existence of a class of desire (“liking”) verbs which categorically disallow *ki* (pp. 81-82). Our data shows that association with this complementizer in fact forms a scale among verbs that express desire and preference, with *roce* (‘wants’) on one end, *meqave* (‘hopes’) on the other end, and *maʔadif* (‘prefers’) somewhere in between (1%-8%-16% occurrence with *ki*).

Similarly, our LLM-based method uncovers more variation than is apparent from existing resources. An example is the verb *megale*, which Zuckermann (2006, 87) translates as ‘discovers’ and classifies as unlikely to occur with *ki*. In our corpus, a substantial 36% of the verb’s occurrence with an overt complementizer are in fact with *ki*. The cluster that is most strongly associated with this complementizer uncovers a second use of the verb, shown in (3). In this use the verb is associated with inanimate subjects and conveys the meaning of ‘reveals’.

- (3) biqu be-yapan megale ki hakol
 visit in-Japan reveals COMP everything
 yaxasi.
 relative
 ‘A visit to Japan reveals that everything is relative.’

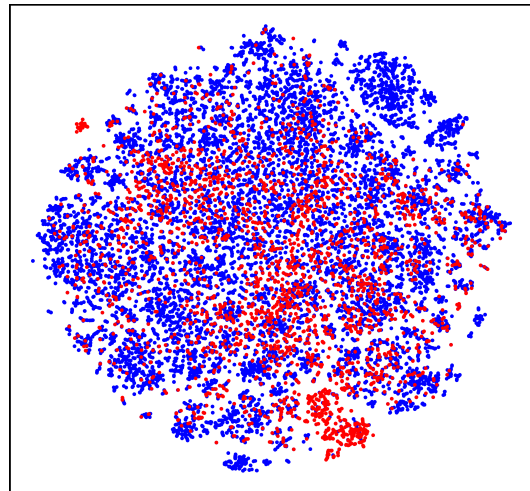


Figure 2: t-SNE plot visualizing complementizer interchangeability for the verb על-י *qala* (‘arose’).

5 Complementizer Propensities: Visualization and Analysis

As described above, in order to visualize the differences in complementizer variability across different types of usages of a given complement-taking verb, we generate t-SNE plots of the contextualized embeddings of the complementizers for a large set of sentences (up to 20K) for each verb.

For example, in Figure 2, we plot 20,000 instances of complementizers which open a subordinate clause for the verb *qala* (in the *qal* conjugation; see על-י in Table 5). Each sentence is represented by a single dot; blue points represent the complementizer *ki*, and red points represent the complementizer *še-*. Crucially, the contextualized embeddings and the t-SNE plot were all computed with a mask over the complementizer. That is, those processes had no knowledge of the label for any given point; the colors were added afterward according to our ground truth labels. Thus, homogeneous sections of a single color on the plot reflect types of sentences which normatively are used with only one or the other of the complementizers.

To be sure, Figure 2 contains far more blue than red, indicating that on the whole, this verb is generally used with the complementizer *ki*. However, red points are interspersed throughout the plot, indicating that the complementizer *še-* is also attested as a practical option for the same types of sentences; i.e., in almost all cases of complement clauses with the verb *qala*, the two complementiz-

ers can be freely interchanged without worrying that the resulting sentence will sound odd or unnatural to present-day Hebrew readers.

However, there are two substantially-sized homogeneous sections on the plot, which indicate specific uses of one or the other of the complementizers. One such section is the red cluster at the bottom of the graph. Inspection of the sentences represented by these points reveals that they all bear inflections of the idiomatic phrase **לא יעלה על הדעת** (*lo yaʕale ʕal ha-daʕat* ‘it is inconceivable’; lit. ‘it would not rise up upon the mind’). The implication, therefore, is that although this verb can generally be used with either complementizer, when it comes to its use within this idiomatic phrase, it is almost exclusively used with the complementizer *še-*.

Conversely, at the top right of the plot, we find a homogeneously blue section. Inspection of these sentences reveals what appears to be a typographic concern: all of the sentences contain a specification of a percentage statistic immediately after the complementizer, written out in digits and a percent sign as in (4). For instance:

- (4) me-ha-duax ʕole ki 83%
 from-the-report arises COMP 83%
 me-ha-maʕasiqim ...
 of-the-employers
 ‘From the report one gleans that 83% of the employers ...’

In these cases, the orthographic distinction between the two complementizers comes into play. Whereas the two-letter complementizer *ki* is written as an independent word, the single letter complementizer *še-* is prefixed in print to the subsequent word. Typing a single Hebrew letter immediately adjacent to a sequence of numbers and the percent sign may result in jumbled text in some text editors (which have problems combining the right to left Hebrew text with numbers, which are written from left to right), or the visual anomaly of a single hanging letter in the text may lead writers to insist on the complementizer *ki* in such situations.

In contrast, the plot for the verb *hoxiʕ* (Figure 3) has no solid homogeneous clusters; rather, the red points are fairly evenly interspersed throughout the plot. This indicates that for this verb, both complementizers are accessible. Even though the use of *ki* is somewhat more frequent with this verb, it can optionally be switched out for *še-* without

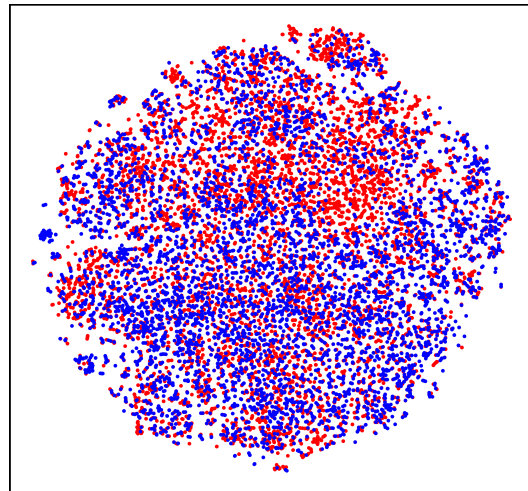


Figure 3: t-SNE plot visualizing complementizer interchangeability for the verb **יכח_הפעיל** *hoxiʕ* (‘proved’).

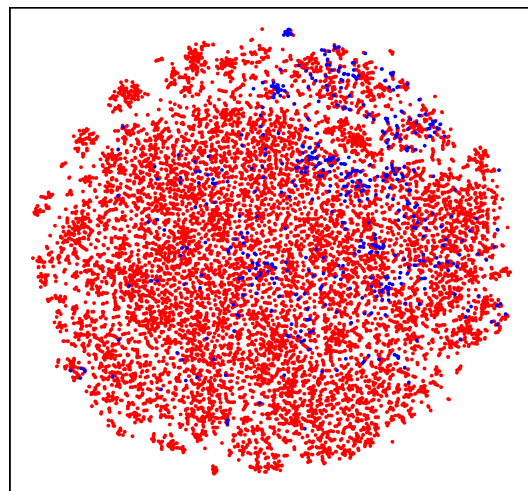


Figure 4: t-SNE plot visualizing complementizer interchangeability for the verb **נגד_הפעיל** *higid* (‘told’).

violating the expectations of native speakers.

A reverse phenomenon emerges from the plot for the verb *higid* (Figure 4), a suppletive form in the paradigm of ‘said, told’. As opposed to the plot for *ʕala*, in which there were a few isolated homogeneous clusters, here the predominant majority of the plot is homogeneously red, reflecting the fact that this verb is used almost exclusively with the complementizer *še-* when it embeds a clause; the complementizer *ki* is generally not a practical option. However, there are a number of areas in the plot where we find blue points interspersed among the red. Inspection of these points reveals that they are sentences with a particular syntactic-semantic profile, exemplified in (5).

- (5) a. nitan lehagid ki avodato šel
 possible to.say COMP work.3MSG of

- hoqusai ...
Hokusai
'It is possible to say that Hokusai's work (had a profound effect on popular themes in woodcut printmaking).'
- b. yeš še-yagidu ki saxqaney
EXIST that-will.say COMP players
fifa ...
FIFA
'Some will say that FIFA players (are not considered real players).'

In these cases, the impersonal nature of the sentence is correlated with the deviation from the verb's general tendency of complementizer selection.

The three t-SNE plots that we have analyzed in this section demonstrate the value of inspecting divergent clusters: the inspection allows researchers to identify the characteristics of clusters which do not follow a verb's general embedding tendency. The ability to do this is of high importance both for NLP purposes and for L2 instruction. In both cases, knowing the general tendency is helpful: in NLP, this allows us to build stronger parsers which are informed by this expectation and which utilize the expectation when disambiguating the sentence tokens; and for L2 learners, knowing the general verb-complementizer pairings can allow them to formulate their sentences in accordance with the expectations of native speakers. However, knowledge of the overarching tendency of a verb is only useful up to a point. Given the clusters we have seen which highlight specialized usages, blind pursuit of the general tendency can lead one astray, whether for L2 language production or for NLP sentence parsing. Instead, the key is to both know the general tendencies of the verbs, and also to know how to identify sentences groups in which those general tendencies do not apply, and may even be flipped. Our t-SNE plots provide an effective method to inspect the usage of the verb overall, and to hone in on examples that constitute specialized usages.

Of course, to the extent possible, we prefer to automatically quantify the extent to which a verb's overall tendency regarding complementizer selection will apply consistently across the full range of its uses, without the need to resort to manual inspection of t-SNE plots. In order to do so, we run an automatic clustering routine on the sentences of a given verb, as described above, and we then calculate the Jensen-Shannon Divergence (JSD) of

each substantially-sized cluster in order to identify divergences.

We utilize this method to calculate corpus-based complementizer statistics for the 100 most frequent complement-taking verbs in contemporary Hebrew. The first part of the table is presented here (Table 3); the full table is presented in Appendix A.⁶ We provide this as a lexical resource for future Hebrew NLP work.

For each verb, we first present the overall statistical tendency towards one or the other of the complementizers, and we note the number of sentences that the statistics were based upon. Additionally, we measure the extent to which the statistical tendency holds true across the corpus, and the extent to which we find specialized usages of the verb in which the tendency is exaggerated or flipped. Regarding many of the verbs, the last three columns contain zeroes, indicating that the overall balance between the two complementizers remains stable for the given verb across the corpus. In contrast, other verbs reveal specialized clusters to greater or lesser degrees. For instance, for the verb *biqueš* ('asked'), the first few columns indicate that in general *še-* is far more likely; at the same time, in 10% of the cases, there is a much stronger affinity to *ki*. On the flipside, regarding *heʔešim* ('accused'), the preference is generally towards *ki*, but 14% of the sentences cluster into groups in which the tendency is flipped toward *še-* instead. Finally, some verbs split in both directions. For instance, for *amar* ('said'), the overall statistics point to balanced usage between the two complementizers, but the divergence columns indicate that, in fact, the usage is often not balanced at all: 21% of the sentences are in clusters that show a specific preference for *ki*, and 16% of the sentences are in clusters that show a preference for *še-*. In other words, in over a third of the corpus, it is not the case that the two complementizers are equally interchangeable, but rather, the varying contexts in which *amar* occurs determine which complementizer is expected.

6 Unusual Complementizers in Hebrew Literature

As explained above in Section 3.3, we propose the use of complementizer clusterings in order

⁶Translations provided for each verb represent its most salient meaning as a clause-embedding predicate; other translations may be more appropriate in specific contexts.

Root	Samples	<i>ki</i>	<i>še-</i>	Cases in Divergent Clusters	Divergence toward <i>ki</i>	Divergence toward <i>še-</i>
אוּחַ פִּיעַל <i>otet</i> ('signal')	3532	57%	43%	0%	0%	0%
אִים פִּיעַל <i>iyem</i> ('threaten')	5330	56%	44%	0%	0%	0%
אִמֵּן הַפְּעִיל <i>heʔemin</i> ('believe')	20000	22%	78%	3%	0%	3%
אָמַר קַל <i>amar</i> ('said')	20000	55%	45%	37%	21%	16%
אָשַׁם הַוּפְּעַל <i>hoʔašam</i> ('was accused')	3175	63%	37%	5%	5%	0%
אָשַׁם הַפְּעִיל <i>heʔešim</i> ('accused')	4328	62%	38%	14%	0%	14%
אִשֶּׁר פִּיעַל <i>išer</i> ('confirmed')	16066	70%	30%	0%	0%	0%
בִּהַר הַפְּעִיל <i>hivhir</i> ('clarified')	20000	64%	36%	4%	4%	0%
בּוֹן הַפְּעִיל <i>hevin</i> ('understood')	20000	17%	83%	0%	0%	0%
בַּחֵן הַפְּעִיל <i>hivxin</i> ('noticed')	5011	40%	60%	0%	0%	0%
בִּטַּח הַפְּעִיל <i>hivtiac</i> ('promised')	20000	34%	66%	1%	0%	1%
בִּקַּשׁ פִּיעַל <i>biqueš</i> ('asked')	20000	16%	84%	10%	10%	0%

Table 3: Complementizer Propensities and Divergencies (Initial part of the table; full table appears in Appendix A)

to identify places in which literary authors deviate from the norm. We analyze all instances of clausal complements within a corpus of novels by the modern Israeli author Amos Oz (see Table 7 in Appendix B for the list of books in this corpus). For each case, we extract the verb which governs the complement, and we then run a K-nearest-neighbor routine to classify the sentence within one of the clusters of that verb (as per the clustering from Section 3.2). We then query this data for cases in which the relevant clusters are highly homogeneous - indicating a preference 95% or higher for one specific complementizer - yet the novelist deliberately chooses the other option. Effectively, in these cases, the novelist subtly undermines the reader's expectations.

Our first example, in (6), exemplifies an unexpected use of *ki* in a story by Oz (Oz, 1976, p. 57):

- (6) biršuta agid la ki
with.her.permission will.say.1SG to.her COMP
lo beit marzeax kan
not tavern here
'With her permission, I will tell her that it is not a tavern here.'

The use of the complementizer *ki* with the verb 'told' is exceedingly rare in general. As we saw above, the one cluster in which this verb is naturally used with *ki* is when the statement is impersonal, with the subject generally unspecified. Yet, the statement in (6) could hardly be more personal; it is phrased in the first person, with a personal plea at the beginning ('please my lady' introduces the sentence we see here). The use of

ki as a complementizer in this context conflicts with the reader's expectations, and characterizes the statement as subtly unusual. And, indeed, in this paragraph, Oz wishes to paint this character - described in the book as an "elderly poet" - as one who interacts with fairly archaic Hebrew expression. In addition to the originally biblical *ki*, in the continuation of the paragraph, this character uses a number of other archaic (Biblical or Talmudic) words and phrases, such as the negative interrogative הֲלֹא *hālō*, as well as אִימָתַי *ʔeymatay* ('when'), and מִי אֲנוּכִי כִי אֲדַע *mī ʔənoḵī kī ʔēdaʕ* ('who am I to know'). The coupling of the verb 'told' with the complementizer *ki*, while exceedingly unusual for contemporary Hebrew, is in fact well-attested in Biblical Hebrew (e.g. Genesis 3:11, Genesis 31:20, Psalms 92:16, and more). Oz's selection of this complementizer is thus clearly deliberate, serving to help characterize the Biblical idiom of the "elderly poet".

The complementizer *še-* is the more general of the two complementizers (no verb exclusively selects *ki*, as can be seen in the table in Appendix A) and is often thought to be a general-purpose complementizer in Hebrew. However, not all uses of *še-* are equally felicitous. Example (7) is highlighted by our procedure as an unexpected use of the complementizer *še-* in Oz's prose (Oz, 1986, p. 168):

- (7) ve-od raciti lehodiafaxa
and-more wanted.1SG to.inform.you
še-me-ha-mixtav še-šalaxta lanu
that-from-the-letter that-sent.2MSG to.us

imxa baxta bi-dmaʕot
 your.mother cried in-tears
 ‘And I also wanted to inform you that your
 mother wept in tears from the letter you
 sent us.’

In general, the preferred complementizer with the verb *hodiʕa* (‘inform’) is *ki* - 74% across the whole corpus. Moreover, this particular sentence is classified as part of a cluster in which the preference for *ki* is far more extreme: over 95%. The cluster includes multiple specimens of the verb with a second-person pronominal suffix, as we find in Oz’s novel. Such phrases are typical of formal and legal documents, which dish out an objective and impersonal ruling; hence the overwhelming preference for *ki*. Yet, Oz’s context is not legal at all. Rather, it is from a letter written by a woman’s second husband, in which he struggles to connect with his step-son, a boy portrayed as unruly and rough, both in character and in his use of language. Oz’s formulation reflects the letter-writer’s struggle in this endeavor. On the one hand, the sentence begins with the highly formal legalese “to inform you” - a phrasing that normally creates a distanced atmosphere. Yet, the unexpected choice of the less-formal complementizer can be seen as an attempt to step back and make the message more accessible to the boy, more personal and more sensitive.

In sum, our method identifies cases where a Hebrew literary master makes a complementizer choice that goes against the grain of how attitudes are usually expressed, inviting further literary analysis to suggest what may have motivated the oddity.

7 Conclusion

This paper demonstrates a language-agnostic method to run a large-scale corpus-based investigation of complementizer variability. We show how this method can be used to isolate cases where authors deviate strikingly from an expected complementizer; such aberrations may well reflect a deliberate literary choice, and invite literary analysis. We apply this method to contemporary Hebrew.

This is the first time that Hebrew complementizers have been investigated from a large-scale corpus-based perspective. Contra the perceived view about Hebrew, we find that the language does have grammatical marking of mood: not in the

verbal morphology, but in its subordinating particles. Moreover, there is not a true subset relation between the uses of the two complementizers; there are verbs that strongly prefer *ki* and allow *še-* only under highly specific contexts. We provide results for complementizer selection regarding the top 100 clause-embedding verbs in contemporary Hebrew. We expect that this first-of-its-kind lexical resource will comprise a helpful resource both for L2 learners, as well as for Hebrew NLP researchers.

Limitations

We demonstrate the ability to identify specialized usages of a verb whose complementizer tendencies differ from the general use of the verb. However, because this method depends on the existence of deviant clusters which highlight the specialized usages, it is inherently limited to usages that are sufficiently well-attested. If a specialized usage only occurs in a few dozen sentences in the corpus, then the exceedingly small cluster that they form will not be sufficient to provide a robust statistic about their complementizer tendencies.

Another limitation inherent in our method is that although we succeed in automatically isolating clusters with specialized complementizer tendencies, we do not currently possess the ability to automatically identify what it is that uniquely characterizes the sentences in that cluster. Rather, once a specialized cluster is identified, it requires human inspection in order to extract the generalized property of the sentences therein.

Ethics Statement

Hebrew data are provided with transliteration and translation as well as in standard Hebrew script in order to increase the accessibility of the paper to native speakers and readers of Hebrew. Transliteration follows the widely used guidelines of the Encyclopedia of Hebrew Language and Linguistics (Khan, 2013).

Acknowledgements

The work of the first author has been funded by the Israel Science Foundation (Grant No. 2617/22) and by the European Union (ERC, MiDRASH, Project No. 101071829), for which we are grateful. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European

Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

- Inge Bartning and Suzanne Schlyter. 2004. [Itinéraires acquisitionnels et stades de développement en français L2](#). *Journal of French Language Studies*, 14(3):281–299.
- Sacha Beniamine, Martin Maiden, and Erich Round. 2020. [Opening the Romance verbal inflection dataset 2.0: A CLDF lexicon](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3027–3035, Marseille, France. European Language Resources Association.
- Hanna Fadida, Alon Itai, and Shuly Wintner. 2014. [A Hebrew verb–complement dictionary](#). *Language Resources & Evaluation*, 48:249–278.
- Anastasia Giannakidou and Alda Mari. 2021. *Truth and Veridicality in Grammar and Thought: Mood, Modality, and Propositional Attitudes*. University of Chicago Press, Chicago.
- Benjamin Kane, Will Gantt, and Aaron Steven White. 2021. [Intensional gaps: Relating veridicality, factivity, doxasticity, bouleticity, and neg-raising](#). In *Proceedings of Semantics and Linguistic Theory (SALT) 31*, pages 570–605.
- Matthew Kanwit and Kimberly L. Geeslin. 2018. [Exploring lexical effects in second language interpretation: The case of mood in Spanish adverbial clauses](#). *Studies in Second Language Acquisition*, 40(3):579–603.
- Geoffrey Khan, editor. 2013. *Encyclopedia of Hebrew Language and Linguistics*. Brill.
- Ron Kuzar. 1993. Nominal clauses in Israeli Hebrew. *Hebrew Linguistics*, 36:71–89. In Hebrew.
- Danit Yshaayahu Levi and Reut Tsarfaty. 2024. [A truly joint neural architecture for segmentation and parsing](#).
- Alda Mari and Paul Portner. 2021. Mood variation with belief predicates: Modal comparison and the raisability of questions. *Glossa: a journal of general linguistics*, 40(1).
- Caterina Mauri and Andrea Sansò. 2016. [The Linguistic Marking of \(Ir\)Realis and Subjunctive](#). In *The Oxford Handbook of Modality and Mood*. Oxford University Press.
- Ellise Moon and Aaron Steven White. 2020. The source of nonfinite temporal interpretation. In *Proceedings of the 50th Annual Meeting of the North East Linguistic Society (NELS)*, pages 11–24.
- Bracha Nir. 2013. [Complementizer](#). In Geoffrey Khan, editor, *Encyclopedia of Hebrew Language and Linguistics*. Brill, Leiden.
- Amos Oz. 1976. *The Hill of Evil Council*. Keter Publishing House.
- Amos Oz. 1986. *Black Box*. Am Oved Publishers.
- Deniz Özyıldız, Ciyang Qing, Floris Roelofsen, Mari-bel Romero, and Wataru Uegaki. 2023. [A crosslinguistic database for combinatorial and semantic properties of attitude predicates](#). In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 65–75, Dubrovnik, Croatia. Association for Computational Linguistics.
- Divna Petkovic and Victor Rabiet. 2016. [La polysémie lexicale et syntaxique de l’alternance modale indicatif/subjonctif – perspectives TAL \(lexical and syntactic polysemy of the modal alternation indicative/subjunctive – NLP perspectives\)](#). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 3 : RECITAL*, pages 80–93, Paris, France. AFCP - ATALA.
- Paul Portner and Aynat Rubinstein. 2020. Desire, belief, and semantic composition: variation in mood selection with desire predicates. *Natural Language Semantics*, pages 343–393.
- Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. 2024. [MRL parsing without tears: The case of Hebrew](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4537–4550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Aaron White and Kyle Rawlins. 2020. [Frequency, acceptability, and selection: A case study of clause-embedding](#). *Glossa: a journal of general linguistics*, 5(1):105.
- Ghil’ad Zuckermann. 2006. Complement clause types in Israeli. In Dixon R. M. W. and Aikhenvald Alexandra Y., editors, *Complementation: A Cross-Linguistic Typology*, Vol. 3. Oxford University Press.

A Appendix: Full Statistics for the Top 100 Verbs

As described above, we calculate corpus-based statistics regarding complementizer propensities for the top 100 complement-taking verbs in the Hebrew language, presented here in a table across the next three pages. For a detailed explanation of the fields in this table, see above, end of Section 5.

Root	Samples	<i>ki</i>	<i>še-</i>	Cases in Divergent Clusters	Divergence toward <i>ki</i>	Divergence toward <i>še-</i>
אוּת_פיעל <i>otet</i> ('signal')	3532	57%	43%	0%	0%	0%
איִם_פיעל <i>iyem</i> ('threaten')	5330	56%	44%	0%	0%	0%
אִמֵן_הפעיל <i>heʔemin</i> ('believe')	20000	22%	78%	3%	0%	3%
אמר_קל <i>amar</i> ('said')	20000	55%	45%	37%	21%	16%
אשם_הופעל <i>hoʔašam</i> ('was accused')	3175	63%	37%	5%	5%	0%
אשם_הפעיל <i>heʔešim</i> ('accused')	4328	62%	38%	14%	0%	14%
אשר_פיעל <i>išer</i> ('confirmed')	16066	70%	30%	0%	0%	0%
בהר_הפעיל <i>hivhir</i> ('clarified')	20000	64%	36%	4%	4%	0%
בון_הפעיל <i>hevin</i> ('understood')	20000	17%	83%	0%	0%	0%
בחן_הפעיל <i>hivxin</i> ('noticed')	5011	40%	60%	0%	0%	0%
בטח_הפעיל <i>hivtiac</i> ('promised')	20000	34%	66%	1%	0%	1%
בקש_פיעל <i>biqueš</i> ('asked')	20000	16%	84%	10%	10%	0%
ברר_התפעל <i>hitbarer</i> ('turned out')	20000	43%	57%	0%	0%	0%
בשר_פיעל <i>biser</i> ('apprised')	3191	43%	57%	0%	0%	0%
גלי_התפעל <i>hitgala</i> ('was discovered')	3665	58%	42%	0%	0%	0%
גלי_פיעל <i>gila</i> ('discovered, revealed')	20000	36%	64%	1%	1%	0%
גרס_קל <i>garas</i> ('held')	10602	64%	36%	2%	2%	0%
דאג_קל <i>daʔag</i> ('ensure')	6472	7%	93%	0%	0%	0%
דגש_הפעיל <i>hidgiš</i> ('emphasized')	20000	74%	26%	0%	0%	0%
דוח_פיעל <i>diveax</i> ('reported')	20000	83%	17%	0%	0%	0%
דרש_קל <i>daraš</i> ('demanded')	15093	37%	63%	3%	0%	3%
ודא_פיעל <i>vide</i> ('confirmed')	14358	17%	83%	1%	0%	1%
זהר_הפעיל <i>hizhir</i> ('warned')	16870	71%	29%	17%	9%	8%
זכר_הפעיל <i>hizkir</i> ('reminded')	19241	44%	56%	0%	0%	0%
זכר_נפעל <i>nizkar</i> ('recalled')	3295	17%	83%	0%	0%	0%
זכר_קל <i>zaxar</i> ('remembered')	20000	21%	79%	24%	0%	24%
חוש_קל <i>xaš</i> ('sensed')	10731	27%	73%	0%	0%	0%
חזי_קל <i>xazi</i> ('predicted')	2929	65%	35%	0%	0%	0%
חכי_פיעל <i>xika</i> ('waited for, anticipated')	2407	1%	99%	0%	0%	0%
חיב_התפעל <i>hitxayev</i> ('obligated oneself')	4573	53%	47%	3%	0%	3%
חלט_הופעל <i>huxlat</i> ('was decided')	2542	62%	38%	0%	0%	0%
חלט_הפעיל <i>hexlit</i> ('decided')	20000	32%	68%	22%	16%	6%
חשב_קל <i>xašav</i> ('thought')	20000	4%	96%	0%	0%	0%
חשד_קל <i>xašad</i> ('suspected')	10822	47%	53%	6%	0%	6%
חשף_קל <i>xasaf</i> ('exposed')	5956	80%	20%	0%	0%	0%
חשש_קל <i>xašaš</i> ('worried')	20000	31%	69%	2%	0%	2%
טען_קל <i>taʔan</i> ('asserted')	20000	65%	35%	3%	3%	0%
ידי_הפעיל <i>hoda</i> ('acknowledged')	20000	41%	59%	8%	5%	3%
ידע_הפעיל <i>hodiʔa</i> ('informed')	20000	74%	26%	8%	2%	6%
ידע_נפעל <i>noda</i> ('became known')	10019	73%	27%	21%	0%	21%
ידע_קל <i>yada</i> ('knew')	20000	12%	88%	0%	0%	0%

Table 4: Table of top 100 verbs (part 1; the table continues on the following pages)

Root	Samples	<i>ki</i>	<i>še-</i>	Cases in Divergent Clusters	Divergence toward <i>ki</i>	Divergence toward <i>še-</i>
יכח_הפעיל <i>hoxiāx</i> ('proved')	20000	33%	67%	0%	0%	0%
יסף_הפעיל <i>hosif</i> ('added')	20000	85%	15%	20%	20%	0%
יצא_קל <i>yatsa</i> ('emerged')	2537	8%	92%	0%	0%	0%
יצע_הפעיל <i>hiciaŷ</i> ('suggested')	11491	32%	68%	22%	0%	22%
כחש_הפעיל <i>hixxiš</i> ('denied')	8314	54%	46%	0%	0%	0%
כרוז_הפעיל <i>hixriz</i> ('announced')	16230	58%	42%	7%	7%	0%
כתב_נפעל <i>nixtav</i> ('was written')	2930	77%	23%	18%	18%	0%
כתב_קל <i>katav</i> ('wrote')	20000	64%	36%	43%	32%	11%
לון_התפעל <i>hitlonen</i> ('complained')	6493	51%	49%	0%	0%	0%
למד_פיעל <i>limed</i> ('taught')	16000	41%	59%	4%	4%	0%
למד_קל <i>lamad</i> ('learned')	8428	27%	73%	0%	0%	0%
מלץ_הפעיל <i>himlic</i> ('recommended')	2628	53%	47%	0%	0%	0%
מסר_נפעל <i>nimsar</i> ('was reported')	7335	84%	16%	4%	0%	4%
מסר_קל <i>masar</i> ('notified, provided statement')	20000	85%	15%	0%	0%	0%
מצא_קל <i>maca</i> ('found')	20000	57%	43%	3%	0%	3%
נגד_הפעיל <i>higid</i> ('told')	20000	3%	97%	0%	0%	0%
נוח_הפעיל <i>heniāx</i> ('assumed')	20000	20%	80%	1%	1%	0%
נסק_הפעיל <i>hisiq</i> ('concluded')	5877	43%	57%	0%	0%	0%
סבר_הפעיל <i>hisbir</i> ('explained')	20000	55%	45%	15%	5%	10%
סבר_התפעל <i>histaber</i> ('turned out')	3167	28%	72%	0%	0%	0%
סבר_קל <i>savar</i> ('opined')	20000	45%	55%	0%	0%	0%
סכם_הפעיל <i>hiskim</i> ('agreed')	14670	35%	65%	9%	0%	9%
סכם_פיעל <i>sikem</i> ('agreed upon')	3684	59%	41%	11%	6%	5%
ספר_פיעל <i>siper</i> ('told')	20000	51%	49%	0%	0%	0%
עדיף_הפעיל <i>heŷedif</i> ('preferred')	3631	8%	92%	0%	0%	0%
עוד_הפעיל <i>heŷid</i> ('testified')	20000	57%	43%	1%	1%	0%
עור_הפעיל <i>heŷir</i> ('commented')	4172	59%	41%	11%	11%	0%
עלי_הפעיל <i>heŷela</i> ('revealed')	13213	67%	33%	13%	0%	13%
עלי_קל <i>ŷala</i> ('arose')	20000	85%	15%	2%	0%	2%
עני_קל <i>ŷana</i> ('replied')	6696	43%	57%	12%	10%	2%
עקש_התפעל <i>hitŷaqeš</i> ('insisted')	5885	32%	68%	5%	0%	5%
ערך_הפעיל <i>heŷerix</i> ('estimated')	20000	70%	30%	0%	0%	0%
פסק_קל <i>pasaq</i> ('ruled')	7979	68%	32%	11%	0%	11%
פרסם_פיעל <i>pirsem</i> ('advertised')	4417	79%	21%	0%	0%	0%
צהר_הפעיל <i>hichir</i> ('declared')	20000	64%	36%	7%	6%	1%
ציין_פועל <i>cuyan</i> ('was mentioned')	3718	86%	14%	0%	0%	0%
ציין_פיעל <i>ciyen</i> ('mentioned')	20000	78%	22%	6%	5%	1%
צפי_פיעל <i>cipa</i> ('expected')	13468	16%	84%	1%	0%	1%
צפי_קל <i>cafa</i> ('predicted')	20000	56%	44%	9%	1%	8%
קבע_קל <i>qava</i> ('decided')	20000	72%	28%	0%	0%	0%
קוי_פיעל <i>qiva</i> ('hoped')	20000	16%	84%	0%	0%	0%
קרי_קל <i>qara</i> ('happened')	4268	2%	98%	0%	0%	0%

Table 5: Table of top 100 verbs, part 2 (continuation from previous page)

Root	Samples	<i>ki</i>	<i>še-</i>	Cases in Divergent Clusters	Divergence toward <i>ki</i>	Divergence toward <i>še-</i>
ראי_הפעיל <i>herʔa</i> ('showed')	20000	51%	49%	15%	7%	8%
ראי_נפעל <i>nirʔa</i> ('seemed')	5712	20%	80%	0%	0%	0%
קל_ראי <i>raʔa</i> ('saw')	20000	28%	72%	1%	0%	1%
רגש_הפעיל <i>hirgiš</i> ('felt')	20000	6%	94%	0%	0%	0%
קל_רמז <i>ramaz</i> ('hinted')	9200	54%	47%	4%	0%	4%
קל_רצי <i>raca</i> ('wanted')	20000	1%	99%	0%	0%	0%
רשם_ההפעל <i>hitrašem</i> ('got the impression')	3715	34%	66%	0%	0%	0%
שוב_הפעיל <i>hešiv</i> ('replied')	11062	67%	33%	12%	12%	0%
קל+לב <i>sam lev</i> ('noticed')	2492	13%	87%	0%	0%	0%
קל_שכח <i>šaxax</i> ('forgot')	6168	12%	88%	0%	0%	0%
שכנע_ההפעל <i>hištaxneʔa</i> ('became convinced')	3221	31%	69%	0%	0%	0%
שכנע_פועל <i>šuxna</i> ('was convinced')	15479	22%	78%	2%	1%	1%
שכנע_פיעל <i>šixneʔa</i> ('convinced')	9204	27%	73%	0%	0%	0%
קל_שמח <i>samax</i> ('was happy')	6309	4%	96%	0%	0%	0%
קל_שמע <i>šama</i> ('heard')	12454	17%	83%	0%	0%	0%
שער_פיעל <i>šiʔer</i> ('assumed')	7254	30%	70%	0%	0%	0%
תרע_הפעיל <i>hitriʔa</i> ('warned')	3852	77%	23%	0%	0%	0%

Table 6: Table of top 100 verbs, part 3 (continuation from previous page)

B Appendix: Corpus of Novels by Amos Oz

Book Name	Word Count	Year of Publication
My Michael	57K	1968
Unto Death	28K	1971
Touch the Water, Touch the Wind	36K	1973
The Hill of Evil Counsel	51K	1976
Black Box	66K	1986
To Know a Woman	62K	1989
Panther in the Basement	30K	1995
Suddenly in the Depth of the Forest	18K	2005
Rhyming Life and Death	25K	2007
A Tale of Love and Darkness	184K	2010

Table 7: List of novels by Amos Oz which we analyzed for complementizer usage. All books were originally penned in Hebrew.

From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations

Erik Henriksson, Amanda Myntti, Saara Hellström, Selcen Erten-Johansson,
Anni Eskelinen, Liina Repo, Veronika Laippala

TurkuNLP, University of Turku

{erik.henriksson, amanda.a.myntti, sherik, selcen.s.erten,
aeske, liina.t.repo, mavela}@utu.fi

Abstract

In corpus linguistics, registers—language varieties suited to different contexts—have traditionally been defined by their situations of use, yet recent studies reveal significant situational variation within registers. Previous quantitative studies, however, have been limited to English, leaving this variation in other languages largely unexplored. To address this gap, we apply a quantitative situational analysis to a large multilingual web register corpus, using large language models (LLMs) to annotate texts in English, Finnish, French, Swedish, and Turkish for 23 situational parameters. Using clustering techniques, we identify six situational text types, such as “Advice”, “Opinion” and “Marketing”, each characterized by distinct situational features. We explore the relationship between these text types and traditional register categories, finding partial alignment, though no register maps perfectly onto a single cluster. These results support the quantitative approach to situational analysis and are consistent with earlier findings for English. Cross-linguistic comparisons show that language accounts for only a small part of situational variation within registers, suggesting registers are situationally similar across languages. This study demonstrates the utility of LLMs in multilingual register analysis and deepens our understanding of situational variation within registers.

1 Introduction

Language varies with context as people adapt their linguistic choices to different situations. *Register variation* refers to the distinct forms of language functionally related to specific situations and communicative purposes (Biber, 1988, 2012; Biber and Conrad, 2019). In the text-linguistic approach to register analysis, the frequent use of linguistic features is assumed to be directly functional for the requirements of the situation (Biber and Egbert, 2023). As a result, text-linguistic register analyses typically start with situational descriptions of

registers (e.g. Biber and Egbert, 2018, Section 2).

Nevertheless, register studies have traditionally focused on the linguistic features characterizing different registers, and much less attention has been given to analyzing the communicative situations in which texts are produced (Biber and Egbert, 2023). Furthermore, existing situational analyses often describe entire registers using the same categorical characteristics—such as medium, setting, communicative purpose, interactivity, and topic (Biber and Conrad, 2019)—which are then used to define register categories. These resulting classes are typically assumed to be situationally discrete.

Some recent studies, however, have provided strong evidence for register-internal situational variation (e.g. Gray, 2015; Biber et al., 2020; Egbert and Gracheva, 2023; Wood, 2024), casting doubts on the possibility of defining registers by any essential situational attributes. This has led to a reconceptualization of registers as *continuous* rather than discrete categories—categories that can be recognized but not strictly defined by linguistic features or situational context (Biber and Egbert, 2023). However, previous research on this variation has been limited to English texts, and its extent in other languages is largely unknown.

In this study, we address this gap by adopting the continuous approach to situational analysis introduced by Biber et al. (2020) and applying it to a large multilingual register-annotated corpus. In this framework, texts are coded for 23 parameters that capture situational variables such as purpose, background assumptions, and source of information, using an ordinal scale from 1 to 6. These annotations allow texts to be viewed within a continuous situational space and grouped into new *situational text types* based on their proximity within this space. The coding scheme was designed to capture the full range of situational factors identified in previous studies, including Biber (1994), Biber and Egbert (2018), and Biber and Conrad (2019).

To annotate a large multilingual corpus for its situational characteristics, we apply a new approach: Instead of manually annotating the texts, we use multilingual large language models (LLMs) for the task. Specifically, we utilize GPT-4o-mini (OpenAI, 2024) and LLaMA 3.1 8B (AI@Meta, 2024) to annotate 8,406 texts from the register-labeled Multilingual CORE corpus (Henriksson et al., 2024) in English, Finnish, French, Swedish, and Turkish. We evaluate the LLM-generated annotations against each other and against a human-annotated sample corpus, demonstrating that the LLMs achieve good accuracy. By integrating these situational annotations with the texts’ existing register and language labels, we conduct multilingual analyses on the relationships between situational context and register, as well as cross-linguistic comparisons of situational variation between registers.

Our analyses show that while registers are partially distinguishable by their situational characteristics, considerable register-internal variation exists across all included languages. These findings align with those reported by Biber et al. (2020) for English. We identify six situational text types—“Advice”, “Information”, “Marketing”, “Personal”, “Opinion”, and “Speech”—each characterized by specific contextual features. These clusters partially align with established register categories but more often reveal situational overlap between registers. Moreover, our cross-linguistic comparisons show that language accounts for only a small portion of the total variance in each register, suggesting that the situational characteristics of registers are generally similar across languages. The code and data used in this study are available at <https://github.com/TurkuNLP/situational-analysis-llm>.

We start by describing the corpus and the LLM-based annotation process, including an evaluation of the LLM annotations against a human-labeled subcorpus. We then explore the situational variation within web registers and identify situational text types that emerge from the data. Next, we examine how these text types align with traditional register categories. Finally, we analyze cross-linguistic situational variation within registers.

2 The register-annotated CORE data

We utilize data from the Multilingual CORE corpus (Henriksson et al., 2024), a large manually register-annotated collection of unrestricted web

	En	Fi	Fr	Sv	Tr	Total
News report	200	200	200	200	200	1,000
Description of a thing or person	200	200	200	200	124	924
Description with intent to sell	105	200	200	200	200	905
Other informational description	200	200	166	94	200	860
Narrative blog	200	200	200	200	52	852
Interactive discussion	200	200	200	118	50	768
Opinion blog	200	163	92	155	58	668
How-to or instructions	159	178	113	95	62	607
Encyclopedia article	64	104	132	200	18	518
Review	169	118	112	49	66	514
Sports report	200	166	66	39	30	501
Spoken	58	30	25	6	32	151
Lyrical	70	13	23	16	16	138
Total	2,025	1,972	1,729	1,572	1,108	8,406

Table 1: Composition of web register dataset.

content spanning 16 languages. The texts in the language subcorpora have been collected using different methodologies at different times.

For the English CORE, data was collected through Google searches targeting highly frequent English 3-grams (Egbert et al., 2015), and annotations were performed via Amazon Mechanical Turk, where each document was labeled by four coders, with a label assigned if at least two coders agreed. The Finnish corpus was sourced from a random sample of the Finnish Internet Parsebank (Luotolahti et al., 2015). The remaining subcorpora were derived from Common Crawl data, following the methodology described in Laippala et al. (2022), including steps such as sampling from various time periods, removing boilerplate content, and deduplication. All register annotations were made by trained experts, using a hierarchical taxonomy with 9 main categories and 16 subcategories.

In this study, we focus on the five largest language datasets in Multilingual CORE: English, Finnish, French, Swedish, and Turkish. We include 13 registers, listed in Table 1, based on the following criteria. Registers must have at least one example from each language and a minimum of 500 examples overall, except for the smaller *Spoken* and *Lyrical* registers, which are included for their situational distinctiveness. Secondly, we treat the *Spoken* and *How-to or instructions* categories as non-hierarchical, as their subcategories are small. For simplicity, we exclude texts with multiple labels or no label at all. Given the class imbalance in the original dataset across categories and languages (see Henriksson et al., 2024, Section 4.4), we randomly sample up to 200 examples from each language-register to balance the data while avoiding excessive downsampling. The resulting dataset is shown in Table 1.

3 Situational annotation using LLMs

Large language models (LLMs) (Brown et al., 2020) have emerged as powerful tools for textual analysis and annotation, with some studies suggesting that their accuracy can even surpass that of human annotators (e.g. Gilardi et al., 2023; Törnberg, 2023; Rathje et al., 2023). In this study, we experiment with two recent models—GPT-4o-mini and Llama 3.1 8B—for the situational coding task. We access GPT-4o-mini via the OpenAI API and deploy Llama 3.1 8B on the Mahti Supercomputer (CSC — IT Center for Science Ltd), using it with PyTorch through the HuggingFace Transformers library. For both models, we set the temperature to 0.01 for consistent responses.

We use the two LLMs to code the 8,406 documents for 23 situational parameters, as listed in Figure 1. For each text, we provide the first 5,000 characters as input to the models, along with the system prompt provided in Appendix A, which instructs the models to rate each parameter from 1 (completely disagree) to 6 (completely agree) based strictly on the given text. We also ask the models to briefly explain each scoring decision, which, in preliminary tests, significantly improved both models’ performance. Both models generated the data in the requested output format without any issues.

We compare the inter-annotator agreement (IAA) of the LLM-generated annotations with each other and with a human-annotated sample consisting of 150 documents across all five languages, annotated by multiple human coders. The human annotators, all experts in the CORE label scheme, were given the parameters and documents without any additional guidance on how to annotate them and without being shown the texts’ register labels. The results of these IAA evaluations are presented in Table 2.

	Kappa	Pearson’s R	Support
GPT4-o-mini vs. Llama 3.1 8B (<i>full data</i>)	0.73	0.76	8,406
GPT4-o-mini vs. Llama 3.1 8B (<i>subset</i>)	0.72	0.75	150
Humans vs. GPT4-o-mini	0.50	0.56	150
Humans vs. Llama 3.1 8B	0.43	0.48	150
Biber et al. (2020)	0.46	0.52	1,002

Table 2: Inter-annotator agreement (IAA) scores.

The agreement between the two LLMs is strong across both the full dataset (8,406 documents) and the human-annotated subset (150 documents), with Cohen’s kappa values of 0.72–0.73 and Pearson correlations of 0.75–0.76. In comparison, the agree-

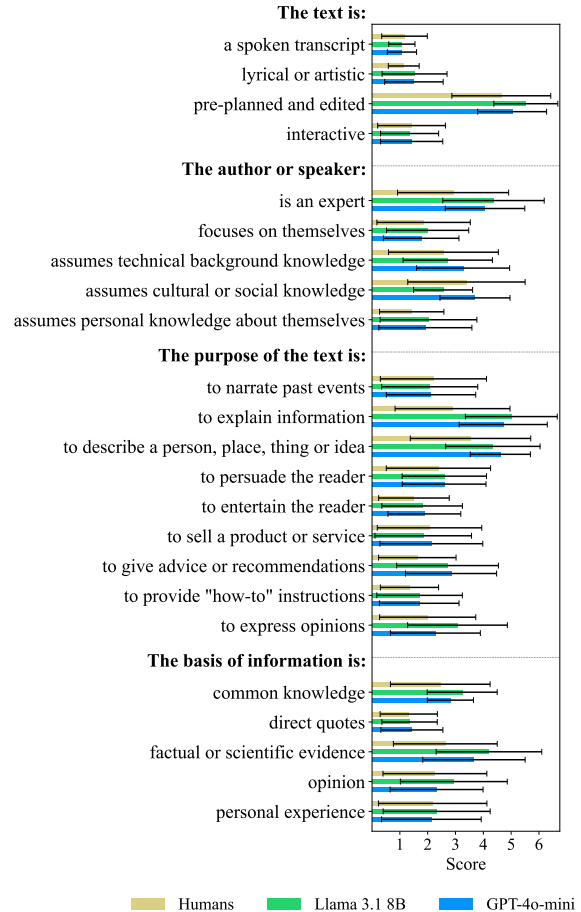


Figure 1: The 23 situational parameters with means and standard deviations, as annotated by GPT-4o-mini, Llama 3.1 8B (full corpus), and human annotators (150 text sample).

ment between the LLMs and human annotators is moderate (Kappa 0.43–0.50, Pearson’s R 0.48–0.56), but these scores are similar to those reported by Biber et al. (2020) for human-made annotations of the same 23 parameters. GPT-4o-mini proves to be slightly more reliable than Llama 3.1 8B when compared to human annotations.

Figure 1 shows the means and standard deviations of the parameter scores from the two LLMs and human annotators. The means of most parameters are relatively close, and all parameters show similar dispersion. This suggests that the moderate-to-strong IAA scores are not simply due, for instance, to the LLMs uniformly selecting the same scores across the dataset. Overall, our results demonstrate that LLMs are well-suited for this annotation task.

In the following sections, we use the dataset of 8,406 texts annotated for the 23 situational parameters, register, and language, to conduct a series of

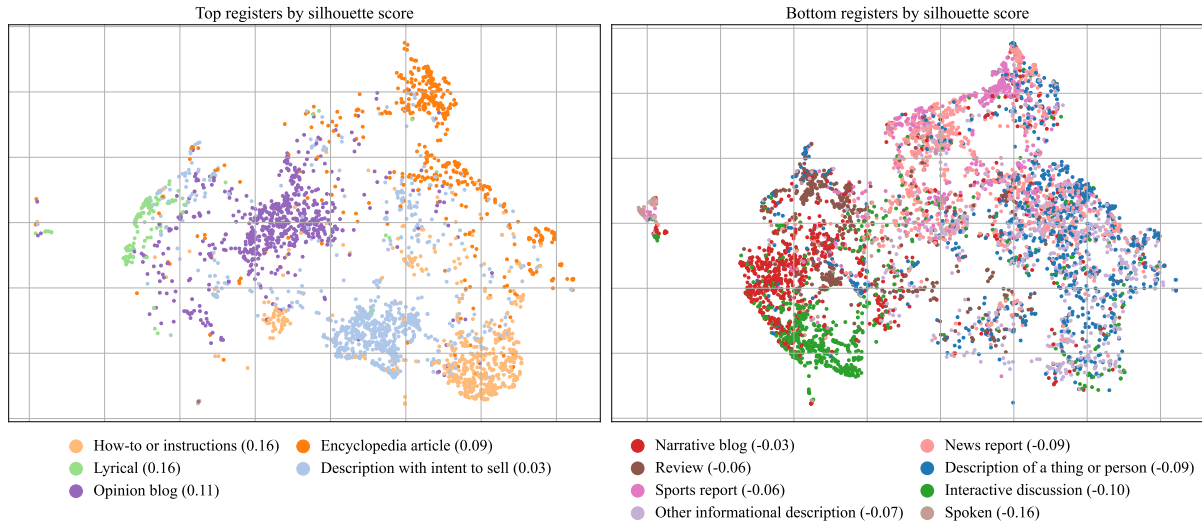


Figure 2: UMAP plots of the registers with the highest (left) and lowest (right) silhouette scores.

analyses. We start by evaluating how well registers are defined by their situational characteristics. Then, we identify distinct *situational text types* directly from the situational data and compare these data-driven categories to traditional register classifications to assess their alignment. Finally, we examine register-internal variation across languages.

4 Quantitative analyses of situation and register

4.1 Registers in a continuous situational space

To begin, we use the LLM-generated annotations to examine how well the register categories in our dataset are situationally defined. To evaluate how well the situational features distinguish each register, we calculate silhouette scores (Shahapure and Nicholas, 2020). This metric measures how similar a text is to the average of other texts within its own register (*cohesion*) compared to the closest instances in other registers (*separation*). Silhouette scores range from -1 to 1, where higher values indicate that instances are better aligned with their own register and more distinct from others.

To visualize the results, we present two UMAP plots (McInnes et al., 2018) in Figure 2. The plot on the left shows the five registers with the highest silhouette scores, while the plot on the right shows the remaining eight with the lowest scores. In both plots, each point represents a text, color-coded by register, with the 2D representation produced by applying UMAP dimensionality reduction to the 23-dimensional situational data.

We observe that the silhouette scores, displayed

next to the register names in Figure 2, are generally very low, ranging from 0.16 to -0.16. The registers *How-to or instructions* and *Lyrical* (0.16) are relatively the most situationally well-defined, showing some separation based on their situational characteristics. *Opinion blog* (0.11) also shows some degree of separation. In contrast, registers like *Interactive discussion*, *Description of a thing or person* (-0.10), and *Spoken* (-0.16) have very low scores, indicating strong overlap with other registers and poor situational definition. In addition to overlap, the low silhouette scores are likely influenced by noise and the presence of numerous outliers in the data. For example, although the *Lyrical* category appears mostly clustered on the left edge of the UMAP plots, there are multiple texts from this register dispersed throughout the plot.

The low score for *Spoken* is particularly notable, given that the situational parameters explicitly include one for spoken transcripts (see Table 2). A manual inspection of situational outliers from this register reveals that many of these outliers are written in formal language (e.g. political speeches, presentations), which the LLMs have interpreted as lacking clear markers of direct speech. This issue likely stems from the LLM prompt not providing clear instructions on how to interpret the parameters. We suspect similar inconsistencies may exist for other parameters as well, and plan to address these in future work.

Visual inspection of the UMAP plots suggests that the registers cluster somewhat better than the low silhouette scores indicate (*Opinion blog*, for

instance, is relatively distinguishable at the center), though this may be partly due to UMAP’s compression and focus on preserving local structure. Finally, we note that the positioning of the texts in the plot generally aligns with intuitive expectations; for instance, *Encyclopedia articles* are mostly grouped on the opposite side from *Lyrical* texts, reflecting their situational and communicative differences.

To summarize, there is some situational delimitation between registers, but the extent of this separation varies, and generally, the situational boundaries between registers are blurry.

4.2 Identifying clusters based on situational parameters

We apply K-means clustering on the LLM-annotated situational data to identify distinct situational categories in our multilingual dataset. This approach offers a new perspective on the contextual distinctions within the web-sourced texts, complementing the similar but English-only analysis presented by (Biber et al., 2020). The resulting *situational text types* represent groups that are maximally similar in their situational characteristics.

The K-means algorithm (MacQueen et al., 1967) partitions the data into clusters by minimizing the sum of squared distances between data points and their respective cluster average points (centroids). Since K-means requires the number of clusters to be specified in advance, evaluating a range of cluster numbers is a necessary preliminary step to determine the optimal number.

We evaluate situational clusters ranging from 3 to 15 using standard metrics. The silhouette score (as already explained in Section 4.1) measures cluster cohesion and separation, with higher scores indicating better-defined clusters. To compare different cluster sizes with the true register labels, we also calculate the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), which measures the agreement between the clusters and the labels. Additionally, we use the Davies-Bouldin Index (Davies and Bouldin, 1979) to assess the average similarity of each cluster to its most similar counterpart, where lower values indicate better separation. Finally, we calculate the within-cluster sum of squares (WCSS) for each cluster size, which helps identify the optimal number of clusters by potentially revealing an “elbow” point where the rate of decrease in WCSS drops, indicating a good balance between cluster number and compactness.

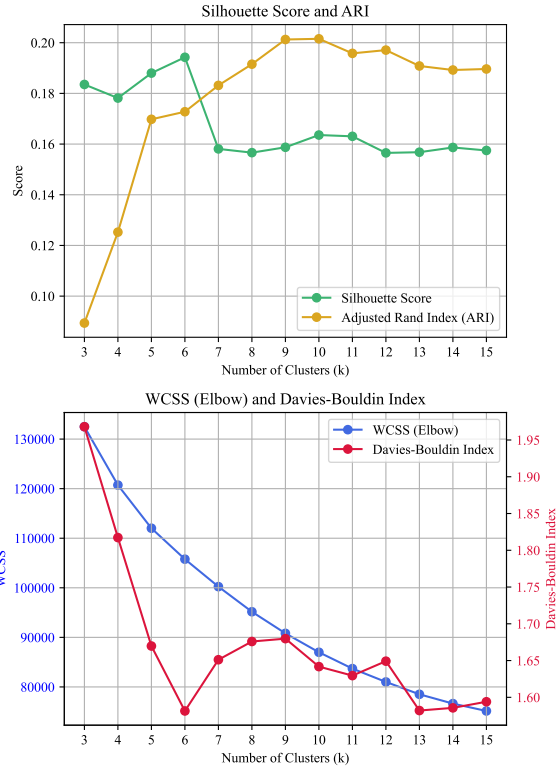


Figure 3: Evaluation of 3 to 15 clusters using silhouette score, ARI, WCSS, and Davies-Bouldin index.

As shown in Figure 3, the average silhouette score peaks at 0.19 with 6 clusters, suggesting optimal cluster definition at this number. At the same point, the Davies-Bouldin Index is also at its most optimal value (1.58), indicating optimal separation. On the other hand, the ARI score is highest with 9 and 10 clusters (0.20), and the WCSS method does not reveal a distinct “elbow” point, as the curve descends smoothly.

Based on these metrics, we select 6 clusters for subsequent analyses, prioritizing cluster cohesion and separation over similarity to the true labels (i.e. high ARI scores), as our goal here is to identify the natural groupings of the texts, independent of their predefined register labels.

We note that all these metrics yield relatively low scores, indicating an overall weak clustering structure for the parameters. This outcome is expected, as the UMAP visualization discussed in Section 4.1 already suggested a lack of clear cluster separation in the data. Furthermore, the low ARI score is unsurprising, given that situational context is only one aspect of what characterizes registers, alongside their linguistic features. Nonetheless, as we show in the next section, the clustering still provides some meaningful differentiation.

Cluster 1: "Marketing" (sil: 0.21, $N = 919$)	M	ΔM
The purpose of the text is to sell a product or service	2.93	3.39
The purpose of the text is to persuade the reader	0.76	1.06
The purpose of the text is to describe a person, place, thing or idea	0.86	0.78
Cluster 2: "Information" (sil: 0.24, $N = 2975$)		
The basis of information is common knowledge	0.51	0.57
The basis of information is factual or scientific evidence	0.85	0.55
The text is pre-planned and edited	0.66	0.44
Cluster 3: "Personal" (sil: 0.17, $N = 1510$)		
The author or speaker focuses on himself/herself	1.90	2.57
The basis of information is personal experience	1.79	2.49
The author or speaker assumes personal knowledge about himself/herself	1.83	2.45
Cluster 4: "Advice" (sil: 0.21, $N = 1117$)		
The purpose of the text is to provide "how-to" instructions	2.57	3.06
The purpose of the text is to give advice or recommendations	1.70	2.14
The author or speaker assumes technical background knowledge	0.61	0.65
Cluster 5: "Speech" (sil: 0.14, $N = 157$)		
The text is a spoken transcript	5.27	5.45
The basis of information is personal experience	1.24	1.94
The basis of information is direct quotes	1.24	1.62
Cluster 6: "Opinion" (sil: 0.12, $N = 1728$)		
The purpose of the text is to express opinions	1.32	1.75
The basis of information is opinion	1.29	1.67
The purpose of the text is to persuade the reader	1.11	1.42

Table 3: Six situational text clusters with silhouette scores (sil.) and number of examples (N). Listed parameters are those with the largest deviations in cluster medians (M) from their global medians (ΔM).

4.3 Interpreting the clusters as situational text types

We now identify the parameters that best characterize each cluster by ranking them based on their typical values within the clusters. Then, we use these rankings to interpret the clusters.

Since the parameter distributions are non-normal (as confirmed by Shapiro-Wilk tests, with p-values < 0.001 in each case), we measure their central tendencies using medians, which are relatively robust against outliers and skewed distributions. To further understand how each cluster stands out relative to the entire dataset, we calculate the deviation of each parameter’s cluster median from the global median for each parameter. This lets us identify which parameters best define each cluster by seeing how much they deviate from the overall trend. The results are shown in Table 3, with descriptive names assigned to each cluster based on their top parameters.

This analysis produces clearly distinguishable situational text types. In “Marketing”, all top parameters relate to the purposes of selling and persuading. “Information” focuses on common, factual, and scientific information. “Personal” centers

on self-reflection and personal knowledge. “Advice” has high scores for instructions, advice, and technical background knowledge, often essential in following instructions. “Speech” includes spoken transcripts, personal experiences, and direct quotes, while “Opinion” is characterized by opinions and persuasion.

The silhouette scores, shown next to the cluster names in Table 3, are low across all clusters (0.12–0.24). This indicates that although the clusters are interpretable based on their top parameters, they are not highly distinct in the situational space. The blurred boundaries between clusters may be partly due to parameters that can be interpreted differently depending on the context. For example, the parameter “The purpose of the text is to persuade the reader” has a high median in both the “Opinion” and “Marketing” clusters, but it serves different functions within these contexts (e.g. arguments in a discussion vs. persuasion with the intent to sell).

4.4 Comparing situational text types and registers

Next, we compare the situational text types, identified in the previous section, with the register categories. The aim is to investigate the mapping between the six data-driven clusters and the 13 human-labeled registers from two perspectives: (1) the composition of each situational text type in terms of registers (*cluster purity*), and (2) the extent to which texts from each register are concentrated within a single situational text type (*register completeness*).

To visualize these alignments, we create a 2D UMAP plot with texts colored by register and overlay a Voronoi diagram (Aurenhammer, 1991), shown in Figure 4. This diagram divides the plot into regions representing each situational text type, with each region containing all points closest to the centroid of the corresponding situational text type.

As Figure 4 shows, there is some alignment between the situational text clusters and the register categories (e.g. *Description with intent to sell* is primarily found within cluster “Marketing”, and *How-to or instructions* is largely in “Advice”). However, no situational text type aligns perfectly with any single register. This imperfect mapping is expected, as (1) the registers are not well-defined situationally, as discussed above in Section 4.1; (2) the clusters were created independently of the register categories by maximizing situational definition;

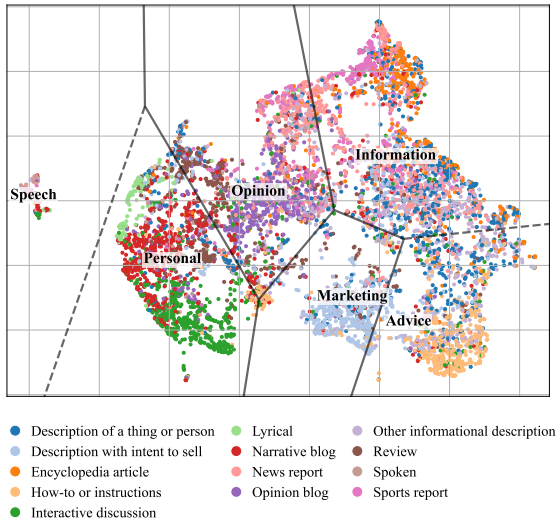


Figure 4: UMAP plot of the six bottom-up situational clusters and 13 registers.

and (3) the number of situational text types differs from the number of registers.

To explore the alignment between clusters and registers in more detail, we present two heatmaps. Figure 5 illustrates *cluster purity*, showing the register composition of each situational text type. Each row represents a situational text type, with columns showing the percentages of registers within each text type.

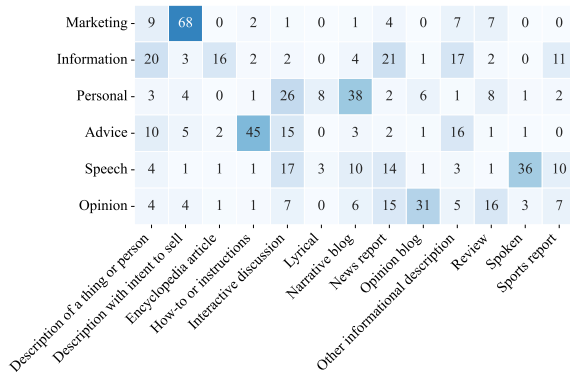


Figure 5: Cluster purity: percentages of registers (columns) in the situational text type clusters (rows).

We find that the register compositions of the clusters generally match the cluster descriptions, though there is significant variation. The “Marketing” cluster, the least variable, includes 68% of texts labeled as *Description with intent to sell*. The “Information” cluster aligns well with informational registers such as *News report* (21%), *Description of a thing or person* (20%), and *Other informational description* (17%). The “Personal” text type

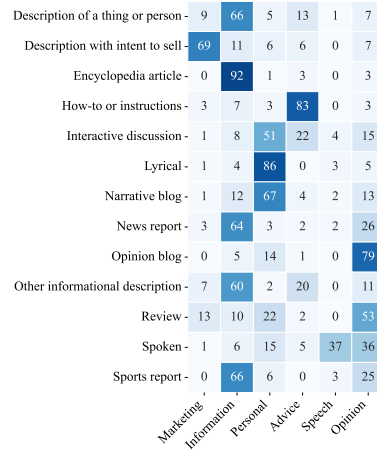


Figure 6: Register completeness: percentages of registers (rows) in the situational clusters (columns).

is primarily composed of *Narrative blogs* (38%) and *Interactive discussions* (26%), where personal matters are often the focus. In the “Advice” cluster, 45% of texts belong to the *How to or instructions* register, followed by *Other informational description* (16%) and *Interactive discussion* (15%), likely providing various forms of advice. The “Speech” cluster includes *Spoken* registers (36%) along with other registers that may contain speech-like elements, such as *Interactive discussion* (17%). Finally, the “Opinion” cluster contains opinionated registers like *Opinion blogs* (31%) and *Reviews* (16%), but also includes *News reports* (15%) and other registers not usually associated with opinion.

The second heatmap (Figure 6) illustrates *register completeness*, showing how registers are distributed across different situational text types. As expected, registers that are more situationally well-defined (see the UMAP plots in Figure 2, Section 4.1) generally map more completely to a single text type. For example, 92% of *Encyclopedia articles* map to “Information”, 86% of *Lyrical* to “Personal”, and 83% of *How to or instructions* to “Advice”. Less well-defined registers, such as *Interactive discussion*, *Review*, and *Spoken*, are more spread across many situational clusters. Notably, the *Spoken* register performs the worst, with texts dispersed across all clusters (1–37%), likely because spoken texts are defined as much by their *purpose* (e.g. expressing opinions) as by the fact that they are spoken.

Interestingly, our multilingual results on the mapping of registers onto situational text types is largely in line with the findings of Biber et al. (2020) for English, which were based on human-

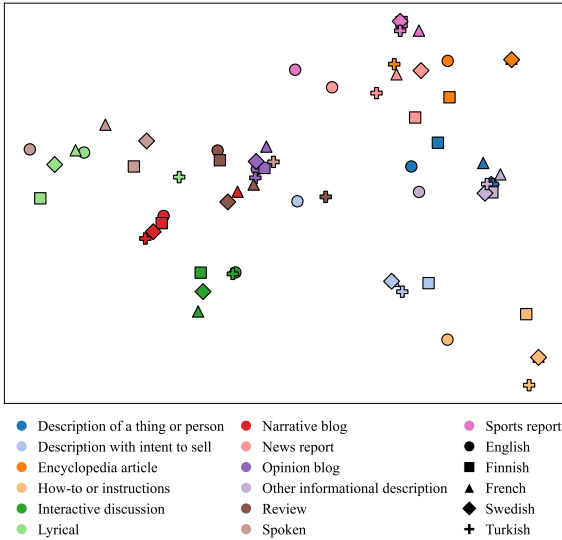


Figure 7: UMAP plot showing situational centroids of registers (colors) across languages (markers).

made annotations. For example, in their analysis, 97% of *Encyclopedia articles* mapped onto a single cluster (vs. our 92%), as did 91% of *Lyrics* (vs. our 86%), and 76% of *How-to* texts (vs. our 83%). This lends further support to the use of LLMs for situational annotation and suggests possible language-independent patterns in register characteristics, though more detailed analysis is needed.

4.5 Cross-linguistic comparisons

Finally, we investigate the similarities of the registers across the included languages—English, Finnish, French, Swedish, and Turkish—in the continuous situational space. As an intuitive way to compare the language-specific registers, we plot their centroids, representing the average position of each register’s data points. We plot the centroids in 2D using UMAP, as shown in Figure 7.

We observe notable variation in how tightly the registers from different languages cluster together in the plot; overall, the registers are not well-separated (consistent with their low silhouette scores; see Section 4.1). The most clearly grouped centroids are those of *Opinion blogs* and *Interactive discussions*, indicating that these registers share similar situational characteristics across languages. Likewise, the *Description of a thing or person* and *Other informational description* registers are also relatively close. *Narrative blogs* are clustered closely in all languages except French. Other registers show more variability, with four

	p	R ²		p	R ²
Description with intent to sell			English vs. Finnish	0.003	0.07
English vs. Finnish	0.001	0.10	French vs. Swedish	0.031	0.06
English vs. Turkish	0.001	0.08	Narrative blog		
English vs. French	0.001	0.07	French vs. Swedish	0.002	0.05
English vs. Swedish	0.001	0.07	News report		
How-to or instructions			English vs. Finnish	0.001	0.05
English vs. Finnish	0.001	0.07	Review		
English vs. French	0.001	0.06	English vs. Turkish	0.002	0.07
English vs. Swedish	0.001	0.06	Swedish vs. Turkish	0.002	0.06
Finnish vs. Turkish	0.001	0.05	Finnish vs. Turkish	0.002	0.05
Lyrical			Spoken		
Finnish vs. Turkish	0.003	0.25	English vs. Turkish	0.005	0.14
Finnish vs. French	0.003	0.21	English vs. Finnish	0.005	0.09
Finnish vs. Swedish	0.016	0.11	French vs. Turkish	0.033	0.06
Swedish vs. Turkish	0.016	0.10	Sports report		
French vs. Turkish	0.010	0.08	English vs. Finnish	0.003	0.08

Table 4: Register comparisons with p-values and R² values showing language-explained variance (where p-values < 0.01 and R² >= 0.05).

of the five languages typically positioned close together, while the remaining language (often English or Turkish) is more distant. In sum, based on a visual examination of the centroids, there is some situational consistency in registers across languages, but the degree of this consistency varies.

To test whether the situational differences between the language-registers are significant, we use PERMANOVA (Permutational Multivariate Analysis of Variance; Anderson 2017), an alternative to ANOVA that does not assume normality, as our data is not normally distributed. We conduct pairwise PERMANOVA tests across all language pairs within each register, applying Bonferroni correction for multiple tests. Additionally, we calculate R² scores to measure the proportion of situational variance explained by language.

Table 4 presents the results for comparisons with p-values < 0.01 and R² >= 0.05. While the tests reveal statistically significant differences across several language pairs, the R² values are generally very low, typically around 0.05 to 0.10, indicating that language explains only a small portion of the total variance in each register. The relatively higher R² values in the *Lyrical* and *Spoken* registers (e.g. 0.25 for Finnish vs. Turkish in *Lyrical*) should be interpreted cautiously due to very small sample sizes (only 6–70 examples per language). The majority of the comparisons (86 of 110), omitted from Table 4, yielded nonsignificant results.

Overall, while there are statistically significant language-specific differences in how registers appear in the situational space, they generally account for only a small part of the total variance. This suggests that most of the situational variance within registers is influenced by factors other than lan-

guage. These factors are worth exploring in future research, though it is beyond the scope of this article.

5 Conclusion

This study explored the situational variation of web registers across multiple languages by utilizing LLM-generated situational annotations alongside manual register labels. Analyzing 8,406 texts in English, Finnish, French, Swedish, and Turkish, we identified six situational text types—such as “Advice” and “Opinion”—that cut across the traditional register categories in the dataset. Our findings indicate that while some registers correspond to specific situational clusters, there is significant variation within registers, supporting the view that registers are better described as situationally continuous rather than discrete. Cross-linguistic comparisons further suggest that situational variance within registers is more influenced by internal variation than by language differences, implying that registers are similarly varied across languages rather than distinctly different. The successful use of LLMs for annotation in this study demonstrates their potential in corpus-linguistic register studies.

Limitations

We excluded texts with multiple or missing register labels for simplicity, which limits the scope of our findings. Future work could explore how such texts are positioned within the situational space using cluster analysis and UMAP plots, offering a new method to analyze hybrid or difficult-to-classify texts (Biber et al., 2020). Another limitation of this study is that we focused solely on situational analysis, without addressing linguistic variation. Given the well-established link between linguistic patterns and situational context, comparing these dimensions presents an interesting direction for future research (Egbert et al., 2024). One approach we plan to explore is analyzing how the situational characteristics of texts align with their positioning in Transformer-based (Vaswani et al., 2017) semantic embedding spaces. Finally, in this study, we could only briefly explore the role of language in accounting for situational variation within registers. In future work, we plan to include more languages and conduct detailed statistical analyses to better understand the situational differences and similarities of registers across languages.

Acknowledgements

We wish to acknowledge FIN-CLARIAH (Common Language Resources and Technology Infrastructure), and CSC – IT Center for Science for computational resources. This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Academy of Finland under grant numbers 358720 and 331297.

References

- AI@Meta. 2024. [Llama 3.1](#).
- Marti J. Anderson. 2017. *Permutational Multivariate Analysis of Variance (PERMANOVA)*, pages 1–15. John Wiley & Sons, Ltd.
- Franz Aurenhammer. 1991. [Voronoi diagrams—a survey of a fundamental geometric data structure](#). *ACM Comput. Surv.*, 23(3):345–405.
- Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.
- Douglas Biber. 1994. An analytical framework for register studies. In Douglas Biber and Edward Finegan, editors, *Sociolinguistic Perspectives on Register*, Oxford studies in sociolinguistics, pages 31–56. Oxford University Press, Oxford.
- Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory*, 8(1):9–37.
- Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.
- Douglas Biber and Jesse Egbert. 2023. [What is a register?: Accounting for linguistic and situational variation within – and outside of – textual varieties](#). *Register Studies*, 5.
- Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. [Reconceptualizing register in a continuous situational space](#). *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020.

- Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.
- Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.
- Jesse Egbert, Douglas Biber, Daniel Keller, and Marianna Gracheva. 2024. Register and the dual nature of functional correspondence: Accounting for text-linguistic variation between registers, within registers, and without registers. *Corpus Linguistics and Linguistic Theory*.
- Jesse Egbert and Marianna Gracheva. 2023. Linguistic variation within registers: Granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory*, 19(1):115–143. Publisher Copyright: © 2022 Walter de Gruyter GmbH, Berlin/Boston.
- Fabrizio Gilardi, Meysam Alizadeh, and Mael Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(30):1–3.
- Bethany Gray. 2015. *Linguistic Variation in Research Articles*. Studies in Corpus Linguistics. John Benjamins Publishing Company.
- Erik Henriksson, Amanda Myntti, Anni Eskelinen, Selcen Erten-Johansson, Saara Hellström, and Veronika Laippala. 2024. Untangling the unrestricted web: Automatic identification of multilingual registers.
- Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.
- Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards universal web parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 211–220, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.
- James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.
- OpenAI. 2024. Gpt-4o-mini.
- Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjeh, Claire Robertson, and Jay J. Van Bavel. 2023. GPT is an effective tool for multilingual psychological text analysis.
- Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748.
- Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Margaret Wood. 2024. Linguistic variation in functional types of statutory law. *Applied Corpus Linguistics*, 4(1):100081.

A Appendix: LLM system prompt

We used the following system prompt for the LLMs for situational coding of texts:

You are an expert in describing multilingual web pages for their situational characteristics. The web pages can be written in any language. There are 23 different situational parameters listed below. Your task is to read the document I give to you and code register characteristics based on the content of the web-scraped text.

For each item, select the number that best represents the text. The scale runs from 1 (Disagree completely) to 6 (Agree completely).

****Guidelines:****

- **Read Carefully**:** Base your coding only on the text's content.
- **Absence of Features**:** Assign a score of 1 if you do not observe any relevant features for a parameter.
- **Objective vs. Subjective Content**:** Score as opinion only if the text clearly expresses personal views or judgments. Otherwise, give very low scores for "opinion" related parameters.

Here are the 23 parameters you will be coding for:

[P1] the text is a spoken transcript [1-6] (explanation)

[P2] the text is lyrical or artistic [1-6] (explanation)

[P3] the text is pre-planned and edited [1-6] (explanation)

[P4] the text is interactive [1-6] (explanation)

[P5] the author or speaker is an expert [1-6] (explanation)

[P6] the author or speaker focuses on himself/herself [1-6] (explanation)

[P7] the author or speaker assumes technical background knowledge [1-6] (explanation)

[P8] the author or speaker assumes cultural or social knowledge [1-6] (explanation)

[P9] the author or speaker assumes personal knowledge about himself/herself [1-6] (explanation)

[P10] the purpose of the text is to narrate past events [1-6] (explanation)

[P11] the purpose of the text is to explain information [1-6] (explanation)

[P12] the purpose of the text is to describe a person, place, thing or idea [1-6] (explanation)

[P13] the purpose of the text is to persuade the reader [1-6] (explanation)

[P14] the purpose of the text is to entertain the reader [1-6] (explanation)

[P15] the purpose of the text is to sell a product or service [1-6] (explanation)

[P16] the purpose of the text is to give advice or recommendations [1-6] (explanation)

[P17] the purpose of the text is to provide 'how-to' instructions [1-6] (explanation)

[P18] the purpose of the text is to express opinions [1-6] (explanation)

[P19] the basis of information is common knowledge [1-6] (explanation)

[P20] The basis of information is direct quotes [1-6] (explanation)

[P21] The basis of information is factual or scientific evidence [1-6] (explanation)

[P22] The basis of information is opinion [1-6] (explanation)

[P23] The basis of information is personal experience [1-6] (explanation)

For each of the 23 points, give a score from 1 to 6 based on the text you read. For each point, explain your given score very briefly, in one short sentence.

In your output, strictly adhere to the following format:

[P1-23] Parameter Name [Your score] (Your explanation)

In the first brackets, write the parameter number [P1 to P23], followed by the parameter name. Then, write your given score in brackets [1-6]. Finally, write your explanation in parentheses ().

Strictly adhere to this output format in all parameter responses. Make sure to fill in all parameters exactly as instructed above.

Evaluating and Adapting Large Language Models to Represent Folktales in Low-Resource Languages

J.A. Meaney, Beatrice Alex and William Lamb

School of Languages, Literatures and Cultures,

University of Edinburgh

jameaney@ed.ac.uk, b.alex@ed.ac.uk, w.lamb@ed.ac.uk

Abstract

Folktales are a rich resource of knowledge about the society and culture of a civilisation. Digital folklore research aims to use automated techniques to better understand these folktales, and it relies on abstract representations of the textual data. Although a number of large language models (LLMs) claim to be able to represent low-resource languages such as Irish and Gaelic, we present two classification tasks to explore how useful these representations are, and three adaptations to improve the performance of these models. We find that adapting the models to work with longer sequences, and continuing pre-training on the domain of folktales improves classification performance, although these findings are tempered by the impressive performance of a baseline SVM with non-contextual features.

1 Introduction

Arguably the biggest development in natural language processing in recent years has been the use of pre-trained large language models (LLMs) such as BERT (Devlin et al., 2018) to transfer the linguistic knowledge from models trained on massive corpora to other tasks, without the need for retraining. These LLMs have unlocked new avenues of investigation for many fields of research, such as digital folklore (Lauer, 2023). Progress has been particularly rapid for high resource languages, such as English, Spanish, Mandarin, as the high threshold of data required by LLMs is easily met by what is available online. However, despite multiple models claiming to represent a selection of low-resource languages, the relative scarcity of data available to train on may have a substantial impact on the model's ability to output a faithful representation of the language. This in turn can restrict digital folklore research to using older technologies, or to working on corpora from higher-resource languages.

The current project is an ambitious contribution to folklore preservation and analysis which initially focused on digitally rendering thousands of hand-transcribed tales in Irish and Gaelic (Sinclair et al., 2022). With this step complete, we now aim to use cutting-edge language technology to uncover cultural and linguistic links between the folklore traditions of Ireland and Scotland.

Our dataset presents two challenges: firstly that it comprises two low-resource languages - Irish and Scottish Gaelic. We are interested in measuring how well LLMs represent the languages in our dataset, using the base version of the model with no adaptations. We also experiment with continuing fine-tuning of these base models with a bilingual corpus of folktale - i.e. domain adaptation. The second challenge is that many of our tales exceed the maximum context window allowed by common LLMs. This leads us to the following research questions:

1. How well do base models work for low-resource languages?
2. Does an increase in maximum sequence length improve performance on this dataset?
3. Does domain adaptation improve performance?

We operationalise these questions by setting up two classification tasks - to predict the type of folktale and the gender of the person who told the folktale.

2 Related Work

Although the field of computational folkloristics is relatively new, predicting the tale type of a folktale has been approached a number of times. Nguyen et al. (2013) classified the tale type of Dutch folktales using the Learning to Rank algorithm. The features to their system included measures of lexical similarity using TFIDF representations and

Jaccard similarity, and subject-verb-object triplets extracted from the text. Their nearest neighbours approach easily outperformed their baselines.

Lindemann et al. (2019) focused not on ATU prediction, but on predicting one of the elements which determines which ATU number is appropriate - locations within the tale. They annotated a corpus of German folktales for location, with moderate inter-annotator agreement. They then used rule-based and statistical approaches to classify the locations. Their best-performing system was a hybrid Naive Bayes system with some rule-based features.

Lô et al. (2020) experimented on a corpus of 742 West African and Western European folktales, all translated into English. Amongst their experiments, they trained classifiers to distinguish between the African and European tales, using a BoW as input to a multi-layer perceptron, as well as word embeddings as input to a long short-term memory (LSTM) system. The much simpler BoW approach performed significantly better than the more complex LSTM system, and the authors speculate that it was the small size of the dataset that led to this.

Eklund et al. (2023) used SVMs to predict the 10 most populated classes of the English-language ATU-annotated tale collections available online (1,518 texts annotated with 182 ATU tale types). They used TFIDF features as input, and reported F-scores ranging between 0.8 and 1.0 for the ten best populated classes. However they did not report any baseline model, so it is difficult to contextualise their achievements.

Related work on Irish and Scottish Gaelic speech and language processing more widely can be found in Lamb and Fransen (In press), including work on handwriting recognition of the folktale data used for this project (Sinclair et al., 2022; Ó Raghallaigh et al., 2022). The current paper is the first study applying computational text classification methods to Irish and Scottish Gaelic folktales.

3 Dataset

The Irish and Gaelic folktales dataset is a collection of 4,692 folk tales collected from two national folklore archives. The Irish language data derive from a 40k page subset of the National Folklore Collection of Ireland’s Main Manuscript Collection, hosted at University College Dublin. This collection comprises over 700k manuscript pages of transcribed folklore interviews made by the Irish Folklore Com-

mission in the mid-20th century. Ó Raghallaigh et al. (2022) describe the digitisation and semi-automatic handwriting recognition (HWR) of our subset of this material, which amounts to 2091 transcription MSS and 3,829,559 words. The average length of the Irish folktales is 1,831.86 words (SD = 1,827.26).

The Gaelic data come from the University of Edinburgh’s School of Scottish Studies Archives. Sinclair et al. (2022) describe how the dataset was created by scanning and semi-automatically recognising 2601 transcriptions and published versions of Scottish Gaelic folklore, amounting to 3,048,348 words. The average length of the Scottish folktales is 5,869 (SD = 12,843).

This dataset includes extensive metadata, such as the type of tale, and the gender of the interviewee who contributed the tale - these two variables are the focus of the below classification tasks detailed in Section 4. The type of tale is determined by the ATU number assigned to it (Uther, 2004). The ATU scheme is a taxonomy used specifically for International Folktales. Each number represents a distinct tale type and was assigned to our tales by professional folklorists. ATU numbers run from 1-2499 and have three main divisions: 1–299 (Animal Tales), 300–1199 (Ordinary Folktales) and 1199–2499 (Jokes, Anecdotes and Formula Tales).

3.0.1 Domain Adaptation Data

There are several aspects of our dataset which likely differ from the data used to train the LLMs we experiment on, this can result in a domain mismatch. As much of our data was collected 30+ years ago, historical language change is a factor, as is the genre of the folktale itself differing from day-to-day speech. For this reason, we experiment with domain adaptation, which continues the pre-training of the LLM in order to help it to learn more about the domain of folktales and style of language used in the past.

Two additional datasets were used for domain adaptation. One is a 400k-word collection of orthographically-normalised Gaelic folktales derived from the Calum Maclean Collection.¹ These are verbatim transcriptions of fieldwork recordings of Gaelic-speaking tradition bearers in the mid-20th-century. The other, known as ‘The Schools’ Collection’,² is a body of folklore taken down by

¹<https://www.calum-maclean-project.celtscot.ed.ac.uk/home/>

²<https://www.duchas.ie/en/info/cbe>

Irish school children from family members and neighbours, also in the mid-20th-century.

4 Methodology

We set up two classification tasks - gender and ATU prediction. There were over 300 ATU types represented in our dataset, so in order to avoid issues of data sparsity, we binned the tale types in our corpus into these four broad categories: ‘Animal’ (ATU 0-299), ‘Magic’ (ATU 300–745), ‘Ordinary (non-Magic)’ (ATU 746-1,199), and ‘Jokes’ (ATU 1,200-2,499), in addition to one for Tales of Magic (aka ‘Fairy Tales’: ATU 300–749). While all of the Irish tales have been labelled with an ATU number, only 451 Gaelic tales were appropriate to label this way; those without an ATU number were omitted from the classifier.

Bin	Gaelic	Irish
Magic	158	1149
Jokes	175	481
Ordinary	96	392
Animal	22	69
Total	452/2601	2091/2091

Table 1: Number of tales per ATU bin in Irish and Gaelic

The second classification task was to predict the storyteller’s gender. The gender of the narrator was strongly skewed towards male in both the Irish data (83.6% male) and the Gaelic data (83.2% male). The gender was reported as ‘unknown’ for 58 of the tales in the dataset and these were omitted from the classifier.

	Irish	Gaelic
Count	1814	2566
Male %	83.6	83.2
Female %	16.4	16.8
Length male	$\mu=1887$ ($\sigma = 1825$)	$\mu=1200$ ($\sigma = 2603$)
Length female	$\mu=1709$ ($\sigma = 1500$)	$\mu=689$ ($\sigma = 905$)

Table 2: Tales by Gender and Length

4.1 Metrics

The F1 score is a commonly used metric for classification tasks, based on precision and recall. However, in light of the imbalanced distribution of labels in both tasks, we opted to report weighted F1

score as the metric (Eq. 1)

$$\text{Weighted F1} = \frac{\sum_{i=1}^k w_i \cdot \text{F1}_i}{\sum_{i=1}^k n_i} \quad (1)$$

where w_i is the weight given by the number of true instances in class c_i , and F1_i is the F1 score for class c_i .

4.2 Baselines

For both tasks, we divided the data into 70% for the training dataset and 15% each for the validation and test set, ensuring that the distribution of languages in the splits was the same as the overall distribution in the dataset.

We set up a dummy baseline for both tasks, which was to select the most frequent label in the training dataset (*Magic* for the ATU task and *Male* for the gender task) and predict this as the correct label for every item in the evaluation. We also built a support vector machine (SVM) in sci-kit learn (Pedregosa et al., 2011), with a simple feature set as input: a count of the tokens which appear in each text, and the term-frequency inverse document frequency (TF-IDF), another count of the tokens with a weighting scheme which gives a higher weight to tokens with lower frequency.

4.3 Transformer Models

We selected three multilingual models that include Irish and/or Gaelic in their training data: mBERT – a version of the BERT model (Devlin et al., 2018) trained on Wikipedia in 104 languages, XLM-RoBERTa (Conneau et al., 2019) which is trained on the Common Crawl in 100 languages and the Language-agnostic BERT Sentence Embedding model (LaBSE), which is trained on 109 languages, with 17 billion monolingual sentences and 6 billion bilingual sentence pairs (Feng et al., 2020). We also included an Irish monolingual model, gaBERT (Barry et al., 2021), which is trained on 171 million tokens of Irish data. To date, no monolingual LLM exists for Gaelic data, so we omit this comparison.

We finetuned these models by using a classification head on top of each model with a loss function of binary cross-entropy for the gender task and cross-entropy for the ATU task. All classifiers were trained for 3 epochs with a batch size of 16 and a learning rate of $2e-5$.

4.4 Adaptations

4.4.1 Length

A known limitation of many Transformer models is that they rely on a self-attention mechanism, whose time- and space-complexity scales quadratically with respect to the sequence length (Keles et al., 2023). For this reason, most use a maximum sequence length of 512 tokens. An alternative to this is the Local, Sparse and Global (LSG) attention introduced by Condevaux and Harispe (2023), which approximates self-attention for sequences up to 4096 tokens. As more than 1,500 of the tales in our corpus exceed the 512 token limit, we adapted all of the models mentioned in the previous section for comparison with their base forms.

4.4.2 Domain

As both of the languages in our dataset are low-resource, and from the specific genre of the folk-tale, we continued finetuning the models on the Schools and Maclean data described in Section 3.0.1. This is referred to as domain-adaptive pre-training (DAPT), and it occurs before adding the classification head. Each model was finetuned for 3 epochs, with a batch size of 32 and a learning rate of $2e-5$.

Model	LSG	DAPT	F1 ATU	F1 Gender
gaBERT	✓	✓	0.69	0.90
SVM	-	-	0.68	0.90
gaBERT	✓	×	0.67	0.89
LaBSE	✓	×	0.65	0.90
LaBSE	×	×	0.62	0.89
gaBERT	×	✓	0.58	0.88
gaBERT	×	×	0.56	0.89
mBERT	✓	✓	0.56	0.87
LaBSE	×	✓	0.54	0.88
LaBSE	✓	✓	0.53	0.84
mBERT	×	✓	0.52	0.88
RoBERTa	✓	✓	0.49	0.82
RoBERTa	✓	×	0.48	0.80
RoBERTa	×	✓	0.43	0.66
RoBERTa	×	×	0.43	0.66
mBERT	✓	×	0.43	0.88
mBERT	×	×	0.40	0.81
Dummy	-	-	0.25	0.66

Table 3: Gender and ATU Classification Results in order of highest F1 on ATU task. Base transformer models are indicated with an × for LSG and DAPT

In terms of our first research question: the base models give very varied performance on this task.

LaBSE performs best for Irish and Gaelic, followed by gaBERT, while mBERT and RoBERTa do not beat the dummy baseline for the gender task, and on the ATU prediction task, they failed to make any correct predictions of the two most infrequent labels - ‘Ordinary’ and ‘Animal’.

Increasing the maximum sequence length improved the performance of the LaBSE and gaBERT models more than domain adaptation did. The combination of LSG and DAPT gave better results than either augmentation on its own for gaBERT, mBERT and RoBERTa, however this combination disimproved the performance of LaBSE over the base model.

Arguably the most interesting result is the strong performance of the SVM. Although the input features were a bag of words and TF-IDF representation, they outperformed almost every model, except the length and domain-adapted gaBERT.

5 Discussion

In order to explore which multilingual LLMs give a faithful representation of two low-resource languages - Irish and Gaelic - we set up two classification tasks on a small folklore dataset, and measured the models’ performance under four conditions: base model, domain-adapted, length-adapted and both adaptations. Of the base models, the language-agnostic LaBSE and the Irish-only gaBERT models performed best. RoBERTa and mBERT did not beat a very simplistic baseline on one of the tasks, and performed poorly at the other. Surprisingly, one of the best results came from an SVM model with non-contextual features. Of the two adaptations we presented, length augmentation improved our results more than domain adaptation, except in the case of LaBSE. The combination of both adaptations gave the best results for gaBERT, mBERT and RoBERTa.

Several interesting points arise from our experiments. Firstly, RoBERTa and mBERT are two of the most well-known transformer models, and their performance on our dataset indicates that they may not represent low-resource languages as well as they claim. Even with adaptations for domain and length, they achieved minimal improvements on the classification tasks.

Secondly, it was not anticipated that gaBERT would perform so well, particularly in light of the fact that it is not trained on any Gaelic data. However, the model’s input data was tokenized with

a SentencePiece tokenizer, and it is possible that there was positive transfer between Irish and Gaelic due to a high number of shared sub-word tokens between the two Celtic languages. It is also notable that domain adaptation did not improve the model’s results much, it was the length augmentation that improve performance. However, as Table 2 indicates, tales told by men make up 83% of the dataset, and with an average length of 1,200 - 1,887 words, it is plausible that allowing the model to process inputs up to 4,096 tokens was beneficial.

Finally, the competitive performance of the SVM is an important takeaway from this paper. Although transformers have yielded state of the art results in NLP for some years now, they have a number of limitations, such as interpretability and energy consumption. On the latter point, the SVM classifier model trained in seconds, compared to 4 minutes for a base or adapted LLM and 30+ minutes for an LSG model. Researchers should consider an SVM as their first choice of model due to its performance, ease of implementation and lower energy footprint.

In order to further understand the performance of our best performing models, we examined the training curves of the SVM and gaBERT models to check for overfitting. We trained the models on subsets of 10%, 30%, 50%, 70% and 100% of the data. After training on each subset, the model is evaluated on the same training data, in order to determine how well it performs on seen data. The validation scores measure the model’s performance on unseen data.

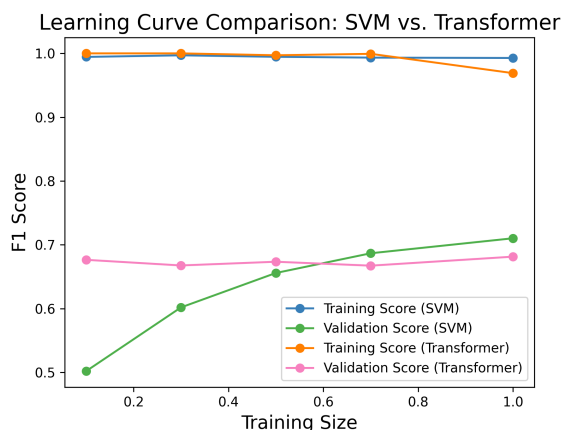


Figure 1: Training Curves and Validation Scores for SVM and gaBERT on the ATU bin Prediction Task

Figure 1 indicates that both the SVM and gaBERT are overfitting to the training data for ATU prediction, as we can see early on that the training

score is high. The disparity between this score and the cross-validation scores for both models also shows that this overfitting means that the models do not generalize as well as hoped to unseen data. Furthermore, the transformer model performs almost as well on 10% of the data as it does on the full dataset.

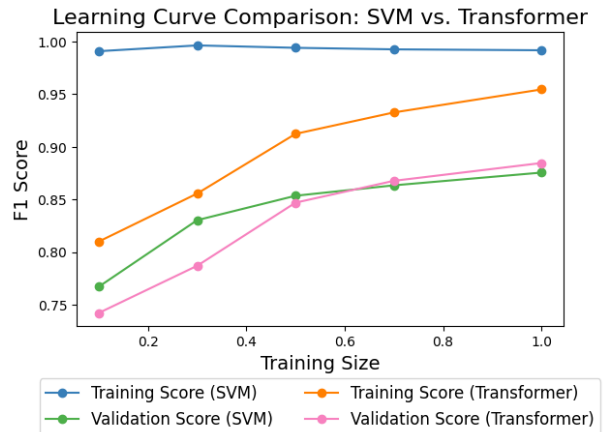


Figure 2: Training Curves and Validation Scores for SVM and gaBERT on the Narrator Gender Prediction Task

Similarly, in the gender task, the SVM overfits the training data, as does the transformer to a lesser extent. However, the transformer’s validation score indicates that it generalises better to unseen data. The performance of the transformer seems to plateau with about 50% of the data, while the SVM could possibly improve performance with some additional data. The use of regularization techniques, or a different kernel may lessen the extent of the overfitting.

6 Conclusion

We present work examining the performance on four LLMs on a classification task, along with three augmentations to the models. The base models do not perform as well as the augmented ones, with mBERT and RoBERTa failing to beat a dummy baseline. The best performing model was trained on Irish data only, was domain-adapted with continued finetuning and used LSG attention to increase the size of its context window. However, our linear baseline - an SVM with bag of words and TF-IDF features gave the second best performance. We conclude that the low-resource setting is challenging for LLMs, and while augmentations do help, classical machine learning models are still a competitive choice.

Limitations

The size of the dataset, at just over 4,600 items, is a limitation. However, it is somewhat characteristic of the low-resource setting. The skewness of the data is a second limitation, there were two labels in the ATU task and one in the gender task that were vastly under-represented, and although we tried to mitigate this through the use of the weighted F1 score, this coupled with the size of the dataset is challenging for large models. Finally, we relied exclusively on extrinsic evaluation, i.e. classification performance. Future work will look at intrinsic measure of LLM’s ability to represent a language, e.g. pseudo-log likelihood.

Ethics Statement

Institutional ethical review for this research was granted by the Ethics Officer of the School of Literatures, Languages and Cultures, University of Edinburgh. No substantial risks are associated with it.

References

- James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J Ó Meachair, and Jennifer Foster. 2021. gabert—an irish language model. *arXiv preprint arXiv:2107.12930*.
- Charles Condevaux and Sébastien Harispe. 2023. Lsg attention: Extrapolation of pretrained transformers to long sequences. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 443–454. Springer.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Johan Eklund, Josh Hagedorn, and Sándor Darányi. 2023. Teaching tale types to a computer: A first experiment with the annotated folktales collection. *Fabula*, 64(1-2):92–106.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2023. On the computational complexity of self-attention. In *International Conference on Algorithmic Learning Theory*, pages 597–619. PMLR.
- Will Lamb and Theodorus Fransen. In press. Irish and scottish gaelic language technology. *Palgrave Handbook of Celtic Languages and Linguistics*.
- Gerhard Lauer. 2023. Computational folktale studies. a very brief history. *Fabula*, 64(1-2):1–6.
- Matthias Lindemann, Stefan Grünewald, and Thierry Declerck. 2019. Annotation and classification of locations in folktales. In *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities, Vienna, Austria, Gerastree Proceedings, GTP*, volume 1.
- Gossa Lô, Victor de Boer, and Chris J van Aart. 2020. Exploring west african folk narrative texts using machine learning. *Information*, 11(5):236.
- Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2013. Folktale classification using learning to rank. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*, pages 195–206. Springer.
- Brian Ó Raghallaigh, Andrea Palandri, and Críostóir Mac Cárthaigh. 2022. Handwritten text recognition (HTR) for Irish-language folklore. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 121–126, Marseille, France. European Language Resources Association.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Mark Sinclair, William Lamb, and Beatrice Alex. 2022. Handwriting recognition for Scottish Gaelic. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 60–70, Marseille, France. European Language Resources Association.
- Hans-Jörg Uther. 2004. *The Types of International Folktales: Animal tales, tales of magic, religious tales, and realistic tales, with an introduction*. 284. Suomalainen Tiedekatemia, Academia Scientiarum Fennica.

Examining Language Modeling Assumptions Using an Annotated Literary Dialect Corpus

Craig Messner

Center for Digital Humanities
Johns Hopkins University
cmessne4@jhu.edu

Tom Lippincott

Center for Digital Humanities
Johns Hopkins University
tom.lippincott@jhu.edu

Abstract

We present a dataset of 19th century American literary orthovariant tokens with a novel layer of human-annotated dialect group tags designed to serve as the basis for computational experiments exploring literarily meaningful orthographic variation. We perform an initial broad set of experiments over this dataset using both token (BERT) and character (CANINE)-level contextual language models. We find indications that the "dialect effect" produced by intentional orthographic variation employs multiple linguistic channels, and that these channels are able to be surfaced to varied degrees given particular language modelling assumptions. Specifically, we find evidence showing that choice of tokenization scheme meaningfully impact the type of orthographic information a model is able to surface.

1 Introduction

Orthographic variation, the deviation from one system of spelling in favor of another, occurs due to a range of intentional and unintentional motivations. Unintentional variation may occur when a writer misspells a word relative to their intended system, or when an optical character recognition system misidentifies a particular character. Intentional deviations are instead used to create a desired political or literary effect (Sebba, 2007). For example, adhering to a system of simplified spelling may signal one's dedication to egalitarian politics, while embedding a literary character's speech in a particular orthographic form may signal an authorial desire to present that character as belong to a particular race, class, region or gender (Ives, 1971) (Jones, 1999).

This latter class of intentional variations proves especially diverse. Supported by the availability of surrounding context and reader-familiar stereotypes of speech, literary orthographic edits are frequently unsystematic ("eye dialect") or not fully beholden to phonetics or morphology (Krapp, 1925).

Instead, the means by which they convey a "dialect effect" is likely multidimensional.

We present a dataset that includes a novel human-annotated layer of dialect family tags designed to support investigations into these varied signalling pathways. We perform an initial set of experiments and discover indications that literary orthographic variation communicates its dialect effect by modifying information along multiple axes: word-level semantics, context-level semantics, and character edits. In the spirit of previous work investigating the phonetic (Agirrezabal et al., 2023), semantic (Rahman et al., 2023) and contextual (Ethayarajh, 2019) information token and character level models capture, we also provide analysis of the literary orthographic understanding of these model types. We additionally offer evidence that character-level models distinguish between intentional literary orthovariants and constructed unintentional variants.

2 Experiments

2.1 Setup

Data. The data for the following experiments consists of 4032 orthovariant tokens paired with their standard forms and sentence-level context, drawn from a 19th century American literature subset of the Project Gutenberg corpus. This corpus is further described in (Messner and Lippincott, 2024). Messner extended the tag set by providing an additional "Dtag" drawn from a set of 31 possibilities, indicating the dialect form ascribed to each observed token.

Messner used the authorially intended subject-position of speaking characters to assign Dtags to tokens. As a result, the Dtag set mostly represents perceived race, nationality, and region. The most populous category (1726 tokens) is the backwoods (BW) tag which combines samples from white-identified northeastern, western and central plains characters. These subcategories are of BW

are often only subtly disjoint; distinguishing them is likely to cause confusion. Other frequent tags include AA (African American: 653), AR (intentionally archaic: 549), GA (Gaelic: 336) and DE (German: 220).

Models. We employ six models for the following experiments. One, fastText-pretrained (Mikolov et al., 2018) is a subword-aware type level embedding model provided by Facebook and trained on CommonCrawl. We use four pretrained token-level contextual models. Two, BERT-large-uncased and BERT-base-uncased (Devlin et al., 2019) use WordPiece tokenization, while CANINE-c and CANINE-s (Clark et al., 2022) are character-level, with the latter utilizing an additional subword loss function during training. Finally, BERT-forced is BERT-base-uncased configured to encode input strings using only single character WordPiece tokens.

2.2 Procedure

Embeddings: the absolute set. We truncate the dataset, keeping only samples that fit the BERT-forced limit of 512 characters, resulting in 3871 observed-standard pairs. For each pair we generate four additional synthetic tokens:

1. **rev:** The standard word in reversed character order. Ex: circus -> sucric
2. **ocr:** A mutated version of the standard word produced using the nlpaug (Ma, 2019) library’s OCR error engine. Ex: circus -> cikcos
3. **swp:** The standard word with a single character swap Ex: circus -> icrcus
4. **rnd:** The standard word with a randomly mutated single character Ex: circun

We collect embeddings for this full token set. For the type-level model, we embed each individual word. For the contextual models, we insert each variant into the context sentence in turn, embed the full sentence, and extract the set of embeddings that represent the target word. For the BERT family of models, we use the last four hidden layers of the model as the embedding values, while for CANINE we use the final hidden layer. If the target word is embedded as more than one subword or character we mean pool the sub-embeddings to generate a final word embedding.

Data augmentation: the relative set. We use these embeddings to produce additional datapoints consisting of the difference between the embedding of a token’s standard form and the embedding of each of the variant forms. Similar to the analogy test of (Mikolov et al., 2013), we use these relative datapoints to investigate a given model’s ability to preserve the intuition that similar types of orthographic transformation should produce similar differences in n -dimensional space.

For each model, we cluster the relative and absolute sets using k -means clustering for each $k \in \{1, \dots, 20\}$.¹

2.3 Evaluation

We use the following measures to evaluate the efficacy of a given k clustering.

Purity. We calculate purity (Manning, 2008) over the clustering of the full relative token set to gain insight into each model’s ability to distinguish between synthetic and observed variants. We also calculate it over the absolute and relative sets of only the observed token to track how well the models cluster embeddings or embedding differences that bear the same Dtag.

Overall accuracy and SO accuracy evaluate a given k clustering’s ability to group token variants from the same datapoint into the same cluster. Overall accuracy is the average percentage of correct groupings of all elements of a datapoint into a single cluster. SO accuracy is the average percentage of correct groupings of only the standard and observed tokens into a single cluster.

Cluster semantic coherency measures the overall semantic similarity of the tokens gathered into a cluster k . We calculate this using the average point-wise cosine similarity of the Word2Vec (Mikolov, 2013) embeddings of each token in a cluster. To support this we train a Word2Vec model on the full corpus using the Gensim (Řehůřek and Sojka, 2010) library.

Cluster Mphone similarity measures the phonetic similarity of the tokens gathered into a cluster k . We calculate this using the average point-wise Levenshtein Distance (LD) of the Metaphone (Philips, 1990) encoded version of each token in the cluster. A lower score indicates that the members of the cluster are more phonetically similar.

¹Code and data for these experiments can be found at <https://github.com/comp-int-hum/orthography-embedding-clustering>

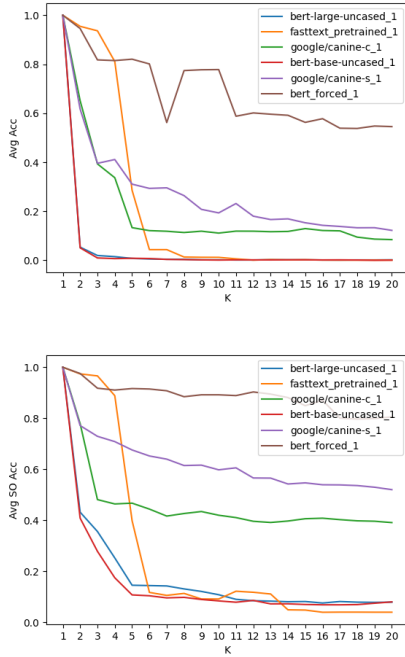


Figure 1: Full absolute set (T), SO absolute set (B) accuracy by k .

3 Results and Discussion

3.1 Evaluating absolute

Only BERT-forced consistently embeds all variants into a similar region. Figure 1 demonstrates that all of the models except for BERT-forced perform uniformly poorly on both overall accuracy across all k , barring the uninformative case where $k < 6$.

The models that perform best on SO accuracy are character-level.

Again barring the uninformative $K < 6$ cases, BERT-forced, CANINE-s and CANINE-c best separate observed-standard pairings from other tokens in their datapoints (Figure 1). Analysis of their shared error reveals that both models perform poorly on a set of high Levenshtein Distance (LD) edit pairs (average LD 2.67). Correspondingly, their shared correct token transformation set has a lower average LD of 1.66. BERT-forced performs better on higher LD transformations, with average correct and error set average LD of 2.2 and 1.9 respectively, implying that BERT-forced preserves difference information beyond character edits.

3.2 Evaluating relative

Of the character and token level models, the CANINE series most distinctly separates con-

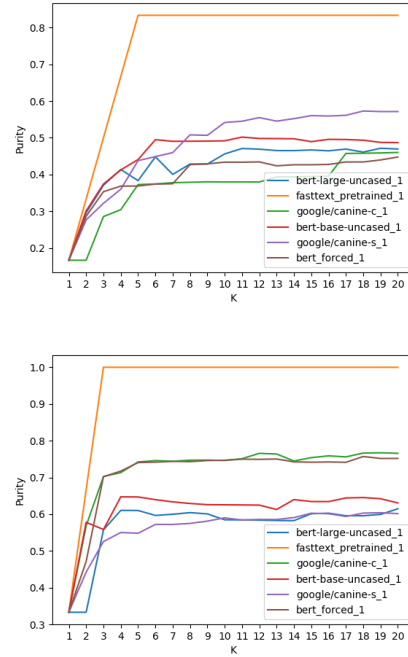


Figure 2: Purity across the full relative set (T) and across non order-swapped tokens (B)

structed and non-constructed variants into clusters. The type-level fastText-pretrained model most accurately separates the variants (Figure 2). However closer examination reveals that it does not separate individual constructed forms, instead grouping them into a single cluster. Notably, character-level models treat order-swapped tokens as functionally similar, while token-level models do not. Removing the rev and swp tokens benefits all models, but overall benefits character-level models the most. Ultimately, this indicates that character-level models preserve information about the distinctions between standard/constructed token differences and standard/observed token differences. It also implies that they rely to a greater degree on the character-edit information stream of the dialect effect to make this determination.

3.3 Evaluation in the light of Dtag and semantic information

High performance on Dtag clustering relies on a mixture of word-semantic, context-semantic and character edit information. As K increases, BERT-base performs best on absolute and CANINE-s on relative (Figure 3). However, both models ultimately only reach purity scores of $\sim .5$, in part at least due to the dominance of the BW tag. Investigating the proportion of individual dtags on

a per-cluster basis at the jointly performant $k=17$ reveals how both models capture a partial mix of these signals.

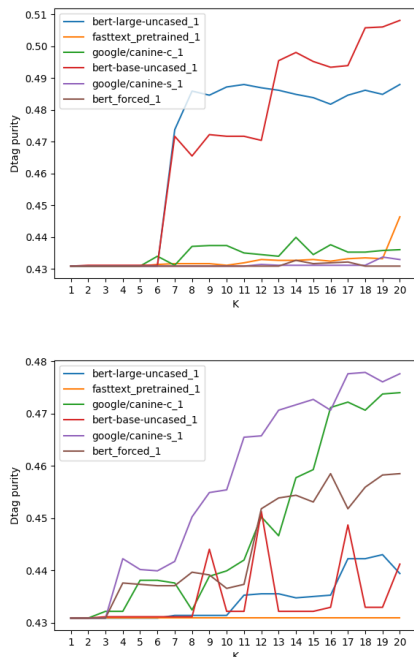


Figure 3: Dtag purity over the obv token embedding (T)
Dtag purity over the std-obv relative set (B)

Certain cluster compositions have potential literary significance. Clusters 3 and 7 of the CANINE-s relative set contain high proportions of both AA (African-American) and WS (White Southern) labeled tokens (Table 1).

K	Count	aa	bw	ws	Mphone
3	160	0.36	0.19	0.24	3.1
7	53	0.59	0.08	0.19	2.5

Table 1: Excerpted Dtag proportions and Mphone similarities of CANINE-s relative set clusters at $K = 17$

Both clusters have low word-level semantic coherence scores (.25 and .27 respectively), consistent with the bulk of of the other clusters at $k = 17$, indicating that this grouping likely does not emerge from word-semantics. This Dtag clustering is particularly striking, as it suggests that period authors took a position on the debate surrounding the origins of southern speech (Bonfiglio, 2010).

The edits shared by WS and AA in these clusters (Table 2) largely impact "r"-related graphemes, demonstrating that this clustering likely occurs due to character edit information. Notably, cluster 3 ranks as less sel-similar than cluster 7 by average

Edit	Standard	Observed
er -> ah	after	aftah
er -> a	rather	ratha
r -> ’	quarters	qua’ters

Table 2: Characteristic edits and examples shared by AA and WS in CANINE-s relative clusters 3 and 7

Mphone LD. Upon inspection, cluster 3 contains a wider variety of "r"-related edits, including r -> y and r -> w. In combination with the somewhat more broad distribution of Dtags found in cluster 3, this implies that these "r" edits are somehow nearer to the sorts of "r" edits characteristic of other Dtag groupings found in the cluster, potentially for contextual reasons.

A similar type of distribution also occurs over GA (Gaelic) tokens. Clusters 8 and 12 contain uniquely high proportions of GA tokens (.27 each) while retaining typically low word-level semantic coherence (.24 and .25). Inspection of the tokens reveals that these clusters collect a variety of edits to the "i" and "e" graphemes. However, unlike the WS and AA clusters examined above, both share similar Mphone LD averages of 3.1 and 3.3 respectively. This may signal that these "i" and "e" transformations are more broadly indicative of a variety of dialect contexts.

Context semantics in part determines accurate literary variant clustering. Notably, the BERT-base absolute set at $k=17$ centralizes clusters around different tags while diffusing WS and GA tokens. For example, cluster 14 has a significantly higher proportion (.33) of DE (German) tagged tokens than any DE-containing cluster found in the CANINE-s relative set.

Edits	Standard	Observed
b -> p	poem	boem
-g	blooming	bloomin
u -> oo	hunters	hoonters
f -> v	falls	valls

Table 3: Characteristic edits and examples of DE tagged tokens in BERT-base absolute cluster 14

The DE tokens in this cluster (Table 3) represent a diverse set of edits, including one (-g in the terminal position) associated with numerous Dtags, including BW (Backwoods) and AA. Given this cluster’s low semantic coherence (.34), a likely conclusion is that this cluster emerges due to the similarity of orthographic contexts in which these tokens appear – say an utterance laden with other

characteristic DE edits.

Low performance on Dtag clustering correlates with high word-semantic cluster coherence in the relative set. For example, BERT-large relative contains multiple clusters with semantic coherence $> .5$, while CANINE-s relative has only one cluster with a score $> .4$. This implies these models favor preserving word-semantic analogical relationships over character edit and context semantics relationships, destabilizing the blend of information needed to successfully cluster over Dtags.

4 Conclusions and Further Work

These experiments offer indications that the dialect effect presented by literary orthographic variation utilizes multiple channels of information: contextual semantics, word semantics and character edits. They also offer evidence that while both contextual token and character level language models can capture all of these aspects, they do so unevenly, justifying further work on the best combination of their information streams.

5 Limitations

The primary limitation of this study emerges from the data. Beyond the inherent limitation of self-restriction to works by 19th century American authors, the coherence of a given observed token and its assigned Dtag is also limited by the inventory of tags chosen. Authors of this period grant their characters multidimensional subject-positions that are reasonably described by but not fully reducible to the granularity of tags like WS and AA. Analysis done in a Dtag-to-cluster direction where the assigned tags are taken as full ground truth limits access to these subtleties.

References

- Manex Agirrezabal, Sidsel Boldsen, and Nora Hollenstein. 2023. [The hidden folk: Linguistic properties encoded in multilingual contextual character representations](#). In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 6–13, Toronto, Canada. Association for Computational Linguistics.
- Thomas Paul Bonfiglio. 2010. *Race and the rise of standard American*, volume 7. Walter de Gruyter.
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. [Canine: Pre-training an efficient tokenization-free encoder for language representation](#). *Transactions of the Association for Computational Linguistics*, 10:73–91.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Sumner Ives. 1971. A theory of literary dialect. *A various language: Perspectives on American dialects*, pages 145–177.
- Gavin Jones. 1999. *Strange talk: The politics of dialect literature in Gilded Age America*. Univ of California Press.
- George Philip Krapp. 1925. *The English Language in America*, volume 1. Century Company, for the Modern language association of America.
- Edward Ma. 2019. Nlp augmentation. <https://github.com/makcedward/nlpaug>.
- Christopher D Manning. 2008. Introduction to information retrieval.
- Craig Messner and Thomas Lippincott. 2024. [Pairing orthographically variant literary words to standard equivalents using neural edit distance models](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 264–269, St. Julians, Malta. Association for Computational Linguistics.
- Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhresch, and Armand Joulin. 2018. [Advances in pre-training distributed word representations](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. [Linguistic regularities in continuous space word representations](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

- Lawrence Philips. 1990. *Hanging on the metaphone*.
- Md Mushfiqur Rahman, Fardin Ahsan Sakib, Fahim Faisal, and Antonios Anastasopoulos. 2023. *To token or not to token: A comparative study of text representations for cross-lingual transfer*. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 67–84, Singapore. Association for Computational Linguistics.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. <http://is.muni.cz/publication/884893/en>.
- Mark Sebba. 2007. *Spelling and society: The culture and politics of orthography around the world*. Cambridge University Press.

Evaluating Language Models in Location Referring Expression Extraction from Early Modern and Contemporary Japanese Texts

Ayuki Katayama[♣] Shohei Higashiyama^{♣,♠,◇} Hiroki Ouchi^{♣,‡,◇} Yusuke Sakai[♣]
Ayano Takeuchi[♡] Ryo Bando[◇] Yuta Hashimoto[†] Toshinobu Ogiso[◇] Taro Watanabe[♣]
♣ NAIST ♠ NICT ◇ NINJAL ♡ NIJL † National Museum of Ethnology ‡ RIKEN
{katayama.ayuki.kc1,hiroki.ouchi,sakai.yusuke.sr9,taro}@is.naist.jp,
shohei.higashiyama@nict.go.jp, takeuchi.ayano@nijl.ac.jp,
yhashimoto@rekihaku.ac.jp, {r-bando,togiso}@ninjal.ac.jp

Abstract

Automatic extraction of geographic information, including Location Referring Expressions (LREs), can aid humanities research in analyzing large collections of historical texts. In this study, we investigated how accurate pretrained Transformer language models (LMs) can extract LREs from historical texts. In particular, we evaluated two representative types of LMs, namely, masked language model and causal language model, using early modern and contemporary Japanese datasets. Our experimental results demonstrated the potential of contemporary LMs for historical texts, but also suggest the need for further model enhancement, such as pretraining on historical texts.

1 Introduction

Historical texts are crucial for a better understanding human and natural history because they record various events and activities of their time. From a *geographic* perspective, historical texts often include Location Referring Expressions (LREs), such as historical place and facility names, along with objects and events related to those locations. As representative examples of such texts, travelogues describe the experiences of the writer in the places they visited, and disaster records describe the affected regions, the scale of the damage, and peoples' situation. Automatic extraction and structuring of such geography-related information by computers can support humanities scholars in analyzing large collections of historical texts.

As a fundamental step for computer-aided geographic text analysis, this study addresses LRE extraction from historical Japanese texts. For an example sentence “名取川渡りて仙台に入る,”¹ an LRE system is required to extract two LREs, “名取川 (Natorigawa)” and “仙台 (Sendai).” Specifically, we investigate the LRE accuracy of Trans-

former (Vaswani et al., 2017) language models (LMs), which have achieved remarkable success in various natural language processing tasks (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020). We focus on two representative types of LMs: Masked Language Model (MLM) and Causal Language Model (CLM).

For model evaluation experiments, we use three datasets: an early modern Japanese travelogue to which we added LRE annotations, Oku no Hosomichi (HOSOMICHI),² early modern Japanese disaster records, the Minna de Honkoku dataset (Hashimoto, 2023) (MINNA), and contemporary Japanese travelogues, the Arukikata Travelogue Dataset (Arukikata Co., Ltd., 2022; Ouchi et al., 2023) (ARUKIKATA). The reasons for using contemporary texts alongside historical texts are twofold: (i) comparing model performance across texts from different eras, and (ii) investigating whether contemporary texts can enhance model performance on historical texts.

Our experiments demonstrated the following results:

- In all settings, an MLM with 3.4M parameters, BERT (Devlin et al., 2019), consistently outperformed a CLM with 7B parameters, Swallow (Fujii et al., 2024).
- The LMs that had been pretrained with contemporary texts achieved high accuracy on the contemporary dataset (F1 scores of up to 0.856 on ARUKIKATA), but yielded low to moderate accuracy on the historical datasets (up to 0.425 on HOSOMICHI and 0.687 on MINNA).
- Models fine-tuned with both contemporary and historical labeled texts achieved the best accuracy for the two historical datasets.

¹The English translation is ‘Crossed the Natori River and entered Sendai.’

²For reproducing our results, we will publish our HOSOMICHI annotation dataset at <https://github.com/naist-nlp/historical-travelogues>.

2 Background and Related Work

LRE extraction, also known as geotagging or toponym recognition, is a special case of named entity recognition (NER) (Nadeau and Sekine, 2007). LRE extraction has often been addressed within the task of geoparsing (Gritta et al., 2020), which aims to estimate the geographic coordinates or geographic database entries that correspond to the locations referenced by LREs.

Resources Previous studies have constructed location-annotated historical corpora and evaluated the performance of machine learning systems in LRE extraction using, for example, English news articles (Coll Ardanuy et al., 2022), English travel writings (Rayson et al., 2017; Sprugnoli et al., 2018), French literary texts (Kogkitsidou and Gambette, 2020), and Chinese historical books (Tang et al., 2024). For Japanese, some researchers have attempted to manually annotate LREs and their geographical coordinates in texts within historical disaster record databases, such as the Database of Materials for the History of Japanese Earthquakes (Kano and Ohmura, 2023) and Minna de Honkoku (Hashimoto, 2023).

System Evaluation Many studies have investigated various methods for recognizing named entities, including locations, in historical texts (Ehrmann et al., 2023). In particular, some recent studies have focused on pretrained Transformer LMs. Labusch et al. (2019) investigated training strategies for BERT suitable for NER on historical German newspaper texts. They showed that a contemporary BERT model achieved the highest accuracy when both pretraining on large unlabeled historical texts and labeled contemporary texts were performed prior to fine-tuning on target labeled historical texts. Tang et al. (2024) evaluated NER accuracy on ancient Chinese historical documents using MLMs pretrained on historical texts, and open and closed contemporary CLMs, with MLMs achieving higher accuracy.

3 Experiments

3.1 Training Scenarios

The purpose of our experiments in this study is to investigate how LMs pretrained with large contemporary texts can be adapted to historical texts. Thus, we employed the following training/evaluation scenarios with three datasets, explained later: (1) fine-tuning on contemporary texts only, (2) fine-tuning

Dataset	Register	#Sentences	#LREs
Arukikata-Train	Travelogue	6,516	3,102
Arukikata-Dev	Travelogue	601	260
Arukikata-Test	Travelogue	5,156	2,166
Minna-Train	Disaster	1,901	9,690
Minna-Test	Disaster	476	2,392
Hosomichi	Travelogue	523	242

Table 1: Dataset categories and statistics.

on historical texts only, (3) fine-tuning on both contemporary and historical texts, and then: (a) evaluating on contemporary texts or (b) evaluating on historical texts. Through these scenarios, we compare the accuracy of an MLM and a CLM.

3.2 Datasets

We curated three datasets in Table 1: a contemporary text dataset (ARUKIKATA) and two early modern text datasets (MINNA and HOSOMICHI).

ARUKIKATA As contemporary Japanese texts, we used the ATD-MCL (Higashiyama et al., 2024), a dataset of travelogues with manually annotated LREs. We treated only LOC-NAME (location name) and FAC-NAME (facility name) mentions as LREs with LOCATION type, and ignored the other mentions. We followed the official train/dev/test split.

MINNA As one of the two historical text datasets, we used the annotation dataset³ from the Minna de Honkoku database (Hashimoto, 2023).⁴ The database comprises records of early modern Japanese disasters from around the 1800s, with manually annotated expressions with “date,” “location,” “damage,” and “person” types. As pre-processing, we divided the single entire document into 50-character segments, and treated each segment as a sentence. Then, we extracted sentences with one or more LREs (i.e., “location” type expressions) and split these sentences into training and test sets at a ratio of 8:2.⁵ Texts in this dataset are typically written in a style where locations and damages at the locations are enumerated, for example, “小石川御門内よりするが臺小川丁筋違御門迄少ゝ破損。”⁶

³<https://github.com/yuta1984/honkoku-data>

⁴https://honkoku.org/index_en.html

⁵We treated LREs across segment boundaries as non-LREs.

⁶LREs are written with underlines. English translation is ‘From inside the Koishikawa Gate to Suruga-dai, Ogawa-cho, and Sujikai Gate—some damage.’

HOSOMICHI As another historical text, we newly created an annotation dataset using the Oku no Hosomichi Wikisource text⁷ (the dataset will be published as mentioned in §1). Oku no Hosomichi is one of the most famous and representative historical Japanese travelogues written by Matsuo Basho in the 1700s. We selected this source because it is a literary text focused on geographic human movement, unlike MINNA, which consists of practical records of geographic events; thus, the two datasets were written in similar periods of time but have different characteristics. Two of the authors manually assigned LOCATION type to the LRE spans within the text. Note that we use this dataset only for evaluation as unseen-domain early modern text because of its limited data size. Texts in Oku no Hosomichi and their English translations can be viewed, for example, on Wikipedia.⁸

3.3 Language Models

We evaluated two types of LMs, MLM and CLM, both of which were pretrained with large contemporary Japanese texts.

MLM We used a character-level Japanese pretrained model⁹ of BERT (Devlin et al., 2019) by fine-tuning it with an additional fully-connected layer for label classification, following the settings in Appendix A. The model is trained to assign one of three labels (B-LOCATION, I-LOCATION, and O) to each character token with the softmax cross-entropy loss.

CLM We used Swallow-7b-hf (Fujii et al., 2024),¹⁰ which has undergone continual pre-training from a Llama 2 (Touvron et al., 2023) model with Japanese language data. We fine-tuned the Swallow model with QLoRA (Detmers et al., 2023) following the settings in Appendix A. Since CLM generates text in an autoregressive manner, we applied the prompt template shown in Figure 1 to each input sentence and fed the filled prompt into the model. The model is then trained to generate the text following “Answer:”. We adopted this simple prompt based on the previous study (Kito

⁷<https://ja.wikisource.org/wiki/%E3%81%8A%E3%81%8F%E3%81%AE%E3%81%BB%E3%81%9D%E9%81%93>

⁸https://en.wikipedia.org/wiki/Oku_no_Hosomichi

⁹<https://huggingface.co/tohoku-nlp/bert-large-japanese-char-v2>

¹⁰<https://huggingface.co/tokyotech-llm/Swallow-7b-hf>

Original Data

```
INPUT_SENTENCE
三日朝ふ廿日迄救民へ にぎり弁当を 廿五 被 下置 御
用 浅草雷門外 深川海辺新田 廿四 御救小屋,
ENTITY_NAME
浅草雷門外
深川
```



Template

```
Sentence: {INPUT_SENTENCE}
Answer: [{"ENTITY_NAME"}; "LOCATION"]
```



Input Data for LLM

```
Sentence: 三日朝ふ廿日迄救民へ にぎり弁当を 廿五
被 下置 御用 浅草雷門外 深川海辺新田 廿四 御救小屋,
Answer: ["浅草雷門外"; "LOCATION"], ["深川";
"LOCATION"],
```

Figure 1: Example of input and output text for Swallow.

et al., 2024), which demonstrated the minimal effect of prompt differences in NER when fine-tuning LLMs.

4 Results and Discussion

Table 2 shows the mean F1 scores of three model runs with different random seeds for each training setting and each evaluation dataset. For each run, the model checkpoint with the best F1 score on the development data was selected.¹¹ We will focus on important aspects in the following sections, and additional discussion is provided in Appendix B.

4.1 On Contemporary Travelogues

On the ARUKIKATA evaluation data, both LMs trained on the ARUKIKATA training data achieved the best F1 scores (0.856 by BERT and 0.797 by Swallow). The models trained on the MINNA training data showed poor accuracy (0.269 and 0.162), and the models trained on the mixed training data did not show any improvement over those trained only on the ARUKIKATA training data. The main reason of these results is the large discrepancy in characteristics between the two datasets; there are differences not only in the

¹¹For the experiments using the MINNA data, we used a random 5% of the training sentences as the development data for training Swallow and the entire training data as the development data for training BERT.

Model	Training data	Evaluation data		
		ARUKIKATA	MINNA	HOSOMICHI
BERT-Large	ARUKIKATA	0.856	0.224	0.345
	MINNA	0.269	0.657	0.361
	ARUKIKATA+ MINNA	0.832	0.687	0.425
Swallow-7b-hf	ARUKIKATA	0.797	0.029	0.244
	MINNA	0.162	0.174	0.257
	ARUKIKATA+ MINNA	0.753	0.267	0.411

Table 2: F1 scores of two LMs on each evaluation data.

era of the texts but also in writing style due to the text register (ARUKIKATA comprising travelogues and MINNA comprising disaster records).

4.2 On Early Modern Disaster Records

On the MINNA evaluation data, we observed the following three findings.

First, both LMs trained on the mixed training data achieved the best F1 scores. These results are somewhat surprising: adding MINNA training data was not effective for evaluation on ARUKIKATA (§4.1), but adding ARUKIKATA training data was effective for evaluation on MINNA. A possible reason is that knowledge of a wide variety of place names may have been useful for MINNA evaluation data; whereas ARUKIKATA data includes a variety of place names from across Japan, MINNA data is biased towards locations around Edo (present-day Tokyo).

Second, the absolute F1 scores for MINNA evaluation data were overall lower than those for ARUKIKATA evaluation data. This would be because the LMs were pretrained on contemporary Japanese texts. LMs pretrained on historical texts can improve the downstream task performance, as demonstrated by Labusch et al. (2019), which is an interesting future direction.

Third, the performance of Swallow (up to 0.267) was substantially lower than that of BERT (up to 0.687). A possible reason is the difference between the training methods, full-parameter tuning for BERT and QLoRA tuning for Swallow. During the QLoRA tuning that we used for Swallow, only a small number of newly added parameters were updated, and the original parameters were fixed. Thus, the model may not be able to fit the training data sufficiently. However, additional evaluation is needed to verify this: tuning BERT with QLoRA. Other possible reasons are the differences in the

approach to the extraction task, namely, classification by BERT and language generation by Swallow, as well as differences in pretraining tasks, namely, MLM and CLM. These could impact the differences in the knowledge acquired during pretraining, as well as the number of examples necessary for downstream task training.

4.3 On Early Modern Travelogue

On the HOSOMICHI evaluation data, both LM achieved close F1 scores when trained on the ARUKIKATA training data and when trained on MINNA training data. Moreover, both LMs trained on the mixed training data achieved the best F1 scores. These results indicate that the two training data were both effective and complementary in the extraction of LREs from the HOSOMICHI data. Probable reasons are as follows. Although the text registers of MINNA and HOSOMICHI are different (disaster records vs a travelogue), the time period of both is relatively close. Although the time period of ARUKIKATA and HOSOMICHI are different (contemporary vs early modern), both travelogue data may be similar in that they include wide range of place names in Japan and describe the writer’s experiences at each location.

Because of the cross-domain scenario, the absolute F1 scores on the HOSOMICHI evaluation data is not high: up to approximately 0.4. Straightforward approaches to improve extraction accuracy for this dataset include pretraining on similar domain texts and fine-tuning with similar domain labeled examples.

4.4 Qualitative Analysis

Table 3 shows LRE examples predicted by the LMs trained on the mixed training data for HOSOMICHI dataset. Although “室の八島 (Muro no Yashima)” is a single LRE, both LMs only rec-

Gold	室の八島 (Muro no Yashima)
BERT	八島 (Yashima)
Swallow	八島 (Yashima)

Table 3: Example LREs predicted by LMs fine-tuned with ARUKIKATA+MINNA data for the sentence “室の八島に詣す (Visiting Muro no Yashima)” in the HOSOMICHI evaluation data.

ognized “八島 (Yashima)” as an LRE. One possible reason is that “の (no)” was misinterpreted as Japanese particle that indicates possession or belonging, leading the LMs to understand it as “Yashima of Muro” or “Muro’s Yashima,” although the entire span is a single phrase. This failure suggests that the LMs lack knowledge about historical place names.

5 Conclusion

This study investigated the extraction accuracy of representative pretrained Japanese LMs using early modern and contemporary LRE datasets. One of main findings from our experiments is the effectiveness of fine-tuning with both contemporary and historical labeled texts. Possible future work includes (i) expanding the evaluation to cover a broader range of eras and registers, and (ii) investigating pretraining strategies using unlabeled historical texts effective for downstream tasks, including LRE extraction and others.

Limitations

In this study, we selected one representative Japanese LM for both MLM and CLM. It is unclear whether similar trends would be observed with other Japanese LMs. Therefore, it is desirable to evaluate a more diverse LMs for a comprehensive analysis in the future. However, considering that most current Japanese LMs are based on the Transformer architecture, we believe that the choice of models is appropriate as a first step in identifying the potential challenges that Japanese LMs may face in extracting LREs from historical Japanese texts.

Ethics Statement

The evaluation datasets present no licensing issues, as ARUKIKATA is under the MIT License, MINNA is under the CC-BY-SA 4.0 License, and HOSOMICHI is sourced from Wikisource under

the same CC-BY-SA 4.0 License. Furthermore, since the original text of “Oku no Hosomichi” is in the public domain, there are no copyright issues related to its distribution. Additionally, the annotation data only involves tagging the original text, which means it does not contain any harmful content in our artifacts.

Acknowledgements

We would like to thank the anonymous reviewers and meta reviewers for their constructive comments. This study was supported by JSPS KAKENHI Grant Number JP23K24904 and NIJAL Joint Resource-use Projects (B) “Geoparsing for Historical Japanese Text.”

References

- Arukikata. Co., Ltd. 2022. Arukikata travelogue dataset. Informatics Research Data Repository, National Institute of Informatics. <https://doi.org/10.32130/idr.18.1>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Mariona Coll Ardanuy, David Beavan, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Katherine McDonough, Federico Nanni, Daniel van Strien, and Daniel C. S. Wilson. 2022. [A dataset for toponym resolution in nineteenth-century english newspapers](#). *Journal of Open Humanities Data*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. [Named entity recognition and classification in historical documents: A survey](#). *ACM Comput. Surv.*, 56(2).
- Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.
- Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. [A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics](#). *Language Resources and Evaluation*, 54:683–712.
- Yuta Hashimoto. 2023. Prototype development of a markup system for historical disaster records. *IPSJ SIG Computers and the Humanities Technical Report*, 2023-CH-131(2):1–6.
- Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. 2024. [Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 513–532, St. Julian’s, Malta. Association for Computational Linguistics.
- Yasuyuki Kano and Junzo Ohmura. 2023. Integration of geographical information into the data of materials for the history of Japanese earthquakes. *IPSJ SIG Computers and the Humanities Technical Report*, 2023-CH-131(3):1–3.
- Taisei Kito, Kohei Makino, Makoto Miwa, and Yutaka Sasaki. 2024. [Koyū hyōgen chūshutsu ni-okeru daikibo gengo model no LoRA fine-tuning no gakushū settei no chōsa](#) (Investigating LoRA fine-tuning training settings of large language models in named entity recognition). In *Proceedings of the 30th Annual Conference of the Association for Natural Language Processing*.
- Eleni Kogkitsidou and Philippe Gambette. 2020. [Normalisation of 16th and 17th century texts in French and geographical named entity recognition](#). In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities ’20, page 28–34, New York, NY, USA. Association for Computing Machinery.
- Kai Labusch, Preußischer Kulturbesitz, Clemens Neudecker, and David Zellhöfer. 2019. BERT for named entity recognition in contemporary and historical German. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 9–11.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.
- Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe. 2023. [Arukikata travelogue dataset](#). arXiv:2305.11444.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).
- Paul Rayson, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. 2017. [A deeply annotated testbed for geographical text analysis: The corpus of lake district writing](#). In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities’17, page 9–15, New York, NY, USA. Association for Computing Machinery.
- Rachele Sprugnoli et al. 2018. Arretium or arezzo? A neural approach to the identification of place names in historical texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics*. CEUR-WS.
- Xuemei Tang, Qi Su, Jun Wang, and Zekun Deng. 2024. [CHisIEC: An information extraction corpus for Ancient Chinese history](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3192–3202, Torino, Italia. ELRA and ICCL.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). arXiv:2307.09288.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz

Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Model Hyperparameters

Table 4 and Table 5 show the hyper-parameters used for BERT-Large and Swallow-7b-hf, respectively.

Hyper-parameter	Value
training epochs	20
batch size	32
learning rate	1e-5
lr scheduler type	linear
warmup ratio	0.1
gradient norm clipping threshold	1.0
optimizer	AdamW

Table 4: The hyper-parameters used for BERT-Large.

Hyper-parameter	Value
training epochs	10
batch size	8
learning rate	5e-5
lr scheduler type	linear
optimizer	paged_adamw_8bit
quant_method	BITS_AND_BYTES
load_in_4bit	True
bnb_4bit_use_double_quant	True
bnb_4bit_quant_type	nf4
bnb_4bit_compute_dtype	float16
lora_alpha	16
lora_dropout	0.1
bottleneck_r	64
torch_dtype	float16

Table 5: The hyper-parameters used for Swallow-7b-hf.

B Additional Experimental Results

Detailed Results for the Main Experiment Table 6 show precision and recall as well as F1 scores of the two LMs in the main experiment, which is shown in Table 2 in §4. For simplicity, the results of the models that achieved the best accuracy among training data settings are shown for each evaluation dataset. We observed that Swallow achieved moderate to high precision (0.533-0.895) for each evaluation dataset, which is not significantly lower than that of BERT (0.367–0.841) and is even higher in two out of three datasets. However, Swallow yielded consistently lower recall than BERT, particularly showing very low recall (0.172) for MINNA. This indicates that Swallow made conservative predictions and that improvements in learning methods or prompts are necessary to enhance coverage.

Effects of Instruction Language in CLM Prompt

We conducted an additional experiment using another prompt for the CLM after the review, based

Model	Train	Eval	P	R	F1
BERT	A	A	0.841	0.872	0.856
	A+M	M	0.662	0.714	0.687
	A+M	H	0.367	0.506	0.425
Swallow	A	A	0.895	0.717	0.797
	A+M	M	0.594	0.172	0.267
	A+M	H	0.533	0.335	0.411

Table 6: Precision (P), Recall (R), and F1 scores of two LMs. “A,” “M,” and “H” represent ARUKIKATA, MINNA, and HOSOMICHI, respectively.

Prompt	Training data	Evaluation data		
		A	M	H
En	A	0.797	0.029	0.244
	M	0.162	0.174	0.257
	A+M	0.753	0.267	0.411
Ja	A	<u>0.304</u>	0.032	<u>0.095</u>
	M	0.164	<u>0.028</u>	0.258
	A+M	<u>0.564</u>	0.272	0.418

Table 7: F1 scores of Swallow with English (En) and Japanese (Ja) prompts. “A,” “M,” and “H” represent ARUKIKATA, MINNA, and HOSOMICHI, respectively.

on a reviewer’s comment that suggested to use (contemporary and early modern) Japanese prompts. Specifically, we used a Japanese prompt, which replaces “Sentence:” and “Answer:” in the original English prompt with “入力文:” and “回答:”, respectively.¹² As show in Table 7, compared to Swallow fine-tuned with the English prompt, the model fine-tuned with the Japanese prompt yielded significantly lower F1 scores in four out of nine settings, which are underlined in Table 7, while it achieved similar F1 scores in the other settings. A possible reason for this degradation is that the backbone model, Swallow-7b-hf, has not been instruction-tuned in the Japanese language. This result suggests the necessity of evaluating more diverse LMs, including instruction-tuned models, as well as investigating LMs’ sensitivity to different prompts.

¹²Because the instruction text of our prompt template is concise, and the model was pretrained in English and contemporary Japanese, we only conducted an additional experiment with the contemporary Japanese prompt. We will investigate the effects of more sophisticated prompts, including those based on early modern Japanese, in the future.

Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction

Woori Jang and Seohyon Jung

KAIST, South Korea

{woori.jang, seohyon.jung}@kaist.ac.kr

Abstract

Literary works present diverse and complex character behaviors, often implicit or intentionally obscured, making character analysis an inherently challenging task. This study explores LLMs' capability to identify and interpret behaviors of artificial beings in 11 award-winning contemporary Korean science fiction short stories. Focusing on artificial beings as a distinct class of characters, rather than on conventional human characters, adds to the multi-layered complexity of analysis. We compared two LLMs, Claude 3.5 Sonnet and GPT-4o, with human experts using a custom eight-label system and a unique agreement metric developed to capture the cognitive intricacies of literary interpretation. Human inter-annotator agreement was around 50%, confirming the subjectivity of literary comprehension. LLMs differed from humans in selected text spans but demonstrated high agreement in label assignment for correctly identified spans. LLMs notably excelled at discerning 'actions' as semantic units rather than isolated grammatical components. This study reaffirms literary interpretation's multifaceted nature while expanding the boundaries of NLP, contributing to discussions about AI's capacity to analyze and interpret creative works.

1 Introduction

Literature has long been a realm where diverse characters interact within narratives, offering deep insights into human nature and the essence of humanity (or non-humanity) across time, cultures, and genres (Piper, 2024; Eder et al., 2010; Frow, 2014). Science Fiction (SF), in particular, presents a more diverse lineage of character types compared to realist novels, featuring various forms of non-human entities — be they animals, aliens, or machines — as active protagonists. This makes SF an ideal genre for exploring literary representations of non-human characters' behaviors.

Identifying and examining character behaviors remains unexplored in both literary studies and text-as-data research. Emulating human reading involves complex cognitive endeavors, including various Natural Language Processing (NLP) tasks like coreference resolution and syntactic structure analysis. Mechanically classifying figurative expressions in literary texts is challenging, but well-designed computational approaches to textual interpretation can lead to new and insightful readings.

We extracted and analyzed the behaviors and cognitive processes of 'artificial beings' in recent Korean SF short stories, using annotations from five human experts and two types of Large Language Models (LLMs). Then we analyzed the agreement rates according to two different calculation methods. We paid special attention to the unique characteristics of the Korean language, which, unlike English, employs a wide variety of endings and auxiliary predicates, often making it impossible to judge the intention and usage by solely looking at a verb's grammatical form. That is, our approach focused on the semantic dimension of action rather than verbs merely as a 'part of speech'.

Given the task's complexity, we utilized state-of-the-art LLMs, known for their proficiency in grasping context and adapting to new tasks. Results show that LLMs can achieve high agreement with human annotators in label assignment for correctly identified spans, demonstrating potential in analyzing intricate literary contexts. However, differences in text span selection highlight ongoing challenges in AI's processing of narrative structures and character identification.

2 Related Works

Character behavior analysis has long held a significant position in traditional literary studies, and with the recent emergence of new approaches integrating computational methods in digital human-

ities, its importance and scope of research have further expanded (Moretti, 2013; Jockers, 2013). A pivotal development in this field was the creation of BookNLP¹, a tool for extracting characters and annotating their attributes from literary texts spanning about 200 years. This annotated dataset has sparked various computational literary studies (Bamman et al., 2019; Sims et al., 2019; Bamman et al., 2020; Soni et al., 2023; Vishnubhotla et al., 2023). Representing this trend, Piper (2024) analyzed the physical actions of characters in English novels to explore how characters’ agency is expressed in literary works.

Concurrently, the NLP field has shown increasing interest in utilizing LLMs for data annotation tasks (Bansal and Sharma, 2023; Ding et al., 2023; He et al., 2024; Alizadeh et al., 2024). This approach is particularly valuable in literary settings where traditional NLP tools struggle. For instance, Hicke and Mimno (2024) have explored using LLMs’ for coreference annotation in literary texts. However, most studies have focused on pre-constructed, extensive literary corpora, with less attention to specific genres or nuanced analyses.

In Korean science fiction, the focus of this study, character studies have been diverse but primarily qualitative (Yoon, 2022; Hong, 2023; Oh, 2023; Lee, 2023). While these studies offer valuable insights into character development and themes, quantitative methodologies or diachronic analyses of specific character types remain largely unexplored.

3 Artificial Being Behavior Dataset and Methodology

3.1 Korean Science Fiction Text

We selected 11 contemporary Korean SF short stories for analysis, annotating full texts — instead of excerpts from a larger set of works — of all 11 stories to ensure each story’s overall theme and the determining characteristics are sufficiently reflected in the data. We investigated all 30 winners of the first to sixth Korean Science Fiction Award (2016-2022), which is currently the most prestigious SF award in Korea, and identified 11 works featuring ‘artificial beings’ as main characters. Detailed information about the 11 stories, including their titles, publication years, lengths, and the names of the artificial being characters, is in Appendix A.

The Korean Science Fiction Award serves as a pertinent object of study in exploring the ‘SF boom’ that swept the Korean literary scene in the late 2010s. Unlike the gradual and robust development of SF in Anglo-American contexts, the Korean literary scene struggled to sustain interest in the genre for decades. Until the 2000s, even major awards aimed at discovering new genre writers often lost momentum after just two or three years. However, this award, launched in 2016 with the slogan “The only domestic SF newcomer literary award, newly born after 10 years,” has prospered, introducing writers who have expanded beyond SF into the broader Korean literary field. It has become an important turning point in Korean SF literature’s evolution and a barometer for contemporary scientific and technological trends. The prevalence of AI and robots-themed works requires our particular attention, for they offer critical insights into perceptions and expectations of artificial beings in modern society.

‘Artificial beings’ here refer to artificially created intelligence or its implemented entity, excluding extraterrestrial life forms or animals, even if depicted as anthropomorphized non-human beings, as well as human to cyborg transformations where the intelligence was not artificially created. Artificial beings in the stories are mainly artificial intelligence, robots, or androids, with varying attributes and behavioral patterns. They exhibit characteristics that parallel human mind and behavior while simultaneously exhibiting unique behaviors and cognitive processes that distinguish their capacity from humans (e.g., entering the cloud, displaying a winking emoticon on the screen, etc.). A thorough categorization and analysis of the vocabulary depicting their behaviors helps explore human-machine boundaries and address ontological questions about future technological societies.

3.2 Data Model Design: Preliminary Experiments

As preliminary experiments before establishing the design of annotation-based research, we conducted several tests to examine the artificial being characters’ behaviors from a lexical perspective. We explored the possibility of automating this process using Python Korean morphological analyzer tools commonly used in the NLP field.

This process revealed that extracting the behaviors of artificial beings from stories without losing their meaning is a delicate task of considerable diffi-

¹<https://github.com/booknlp/booknlp>

Verb Morpheme	Conjugation Examples
돌리다 dollida	화제를 돌리다 (change the subject), 숨을 돌리다 (catch one's breath), 시선을 돌리다 (avert one's gaze), 마음을 돌리다 (change one's mind), 세탁기를 돌리다 (run the washing machine), 문고리를 돌리다 (turn the doorknob)
보다 boda	바라보다 (look at), 생각해 보다 (think about), 장을 보다 (go grocery shopping), 떠보다 (test the waters), 잘못이라고 보다 (consider it a mistake), 피를 보다 (suffer harm)
하다 hada	이야기를 하다 (have a conversation), 준비를 하다 (prepare), 각오를 하다 (be determined), 인사하다 (greet), 후회하다 (regret), 목도리를 하다 (wear a scarf)

Table 1: Examples of Korean phrases where the same verb root is used but has completely different meanings in context.

Dataset	Description	Count
Number of sentences	11 short stories	7,289
Human annotation	5 annotators	9,515
LLM annotation	2 models * 2 versions each	8,575

Table 2: Overview of the dataset.

culty. Table 1 shows examples of verbs used in this paper’s Korean SF short stories corpus that have the same morpheme but completely different meanings. In Korean, it is very common for the same verb form to exhibit semantic diversity depending on the object that the verb governs or the verb’s conjugation pattern. For this reason, it is very difficult to accurately grasp what action a word refers to in the stories using morphological analyzer tools that isolate only the smallest units of meaning. As a result, even if the extracted verbs are categorized, the accuracy is very low. The task’s purpose of extracting only the actions of specific characters in the narrative, coupled with the nature of literary texts where meaning changes significantly depending on the context, further complicates the analysis. Given these factors, we concluded that accurate analysis is difficult with existing NLP tools and designed the annotation work described below.

3.3 Human annotation

Label design and tools: We designed an annotation task where human annotators read all 11 stories from beginning to end, as they 1) mark lexical spans that represent the actions and cognitive processes of artificial being characters, and 2) attach labels to categories they believe these words belong to. While this method is time-consuming and challenging, it allows for a comprehensive un-

derstanding of character behavior patterns without missing the uncertainties and ambiguities that arise in the process of reading fiction. Crucially, this high-context dataset can serve as a foundational resource for training and evaluating LLMs, potentially leading to the development of more sophisticated AI research tools capable of nuanced literary interpretation.

The labels were primarily based on the word supersense tagger (Ciaramita and Altun, 2006) utilized in BookNLP, a tool frequently used in character behavior research. However, as the categories in previous studies were mainly composed of words used to describe human behavior, we redesigned the labels to better reflect the specificity of artificial being characters. Finally, we established 7 labels and 1 Miscellaneous category (to be used when a word is judged not to belong to any other category):

- **Communication.** The exchange of information or ideas between characters.
가쁜 숨을 고르고 서 있는 노인에게 안드로이드가 [말한다says].
나는 어머니의 다음 [말을 기다렸다waited for her next words].
- **Sensory act.** The action or process involving the use of sensory systems, including sensory-based interaction with the environment.
안드로이드는 2층 바닥에서 올라오는 입김을 [감지하고는sensed] 에스컬레이터를 힘겹게 걸어 올라갔다.
그리고 기계 팔을 돌려 에이브를 [바라보았다gazed at].
- **Motion.** Physical movement or change in position, including static states.
[간식도 만들어야prepare snacks] 하고 [장도 봐야do grocery shopping] 하며 화장실 번기도 [뒹아야clean] 한다.
나는 경찰의 안내에 따라 법정에 [들어섰다entered].
- **Body change.** Fundamental alterations in the physical or mental state of a character.
아이들이 주는 간식을 거절하지 못하고 먹다가 [고장이 났던has been broken] ... 몸이 부서지면서도 나를 지켜주던 그 로봇은 이제 없는 거야.
월래 24시간 깨어 있어야 하는 루트는 최소한의 감지 시스템만을 켜둔 채로 [절전모드에 들어갔다went into power-saving mode].
- **Emotion.** Subjective feelings or affective states experienced by a character.
영혼 없이 태어난 아이들을 버리지도 못하면서 그들의 신체 기능이 정지될 때마다 [괴로워했고suffered], 그러면서도 계속 해서 [희망을 품고hold on to hope] 다음 번 태아를 배양했다.
어머니를 처음 만났을 때, 나는 [꿈꾸는 기분이었다felt like (I) was in a dream].
- **Cognition.** Mental processes involved in acquiring knowledge and understanding.
그것이 라디오에서 들었던 총이라는 것을 슬라이드가 당겨지는 순간 [깨달았다realized].
안드로이드는 지구로부터 점점 멀어지기 시작해 ... 하나의 촛불처럼 보이는 우주선의 모습을 [상상한다imagines].
- **Judgement.** The process of forming opinions, making decisions, or drawing conclusions.

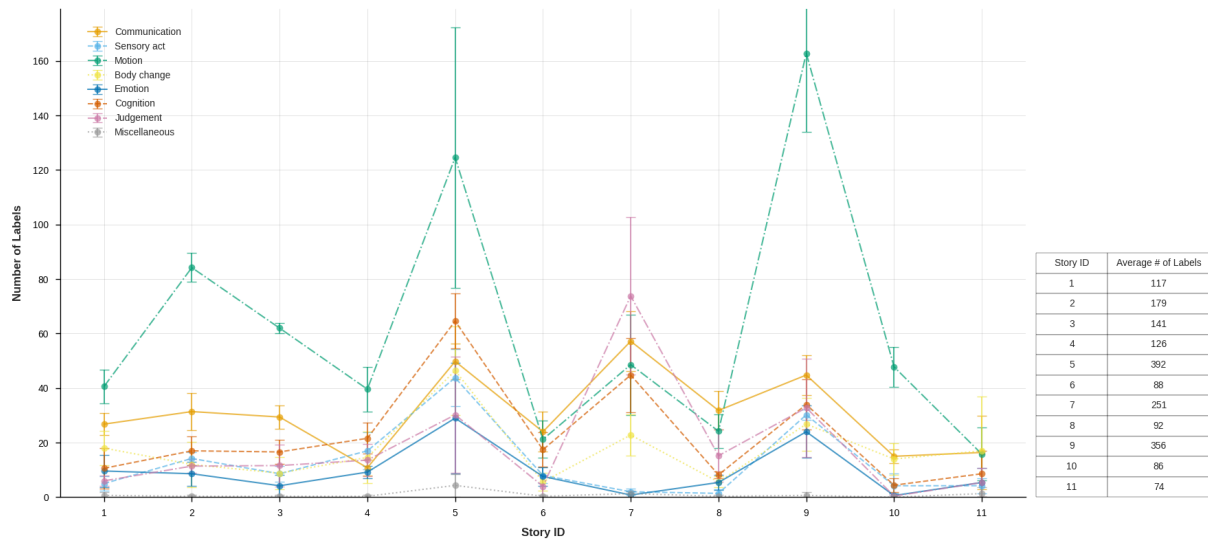


Figure 1: Distribution of labels per story. The numbers from 5 annotators were averaged, and the standard deviation is indicated by shading. There is a variance in the average total number of labels per work, which is shown in the table on the right.

따라 갈 지표가 사라지자 이 큰 건물 안을 돌아봐야 할지, 밖으로 나가야 할지 [판단하지 못했다couldn't decide]. 그러나 자신이 인간을 도울 수 있는 더 큰 힘을 가지고 있다고 [생각하는데considers] '로봇일 뿐'이라니?

Annotators: The annotators ($n=5$)² all experienced in Korean literature, worked independently and did not discuss with each other or change their annotations to match others' annotations to ensure unbiased results. Internal consistency within each annotator's work was prioritized over attempting to establish universally "correct" answers. This approach acknowledges that there may not be a single, objectively correct label for each expression in the story. Instead, annotators were instructed to assign labels to the closest category based on their interpretation of the characteristics, intentions, and context of the artificial being characters portrayed in the narrative. To take into account the unique characteristics of Korean vocabulary, annotators marked minimum spans capturing complete meanings of actions and cognitive processes, often spanning multiple words.

Annotators were provided with full texts of the stories via our annotation tool, as opposed to being given one sentence at a time. They were tasked to span-mark all behavioral vocabulary of artificial being characters and attach single-choice labels from a drop-down menu format. Consequently, the

²Among the five annotators, four were females and one was male. Their ages ranged from the 20s to 30s, and all were native Korean speakers. Two of the annotators held doctoral degrees in literature.

spans of behavioral vocabulary entities marked by each annotator differed, even before considering label differences. The task was far from being an obvious or objective one, yielding many interesting cases with uncertainties or ambiguities.

The number of labels varied depending on the different prominence and characteristics of artificial beings in the stories. Figure 1 shows the average number of labels from five annotators, with standard deviation indicated by error bars. The deviation range illustrates that there was considerable inter-annotator variability, and the dominant labels also varied depending on the narrative.

For instance, Story 5 ("Five Stages of Independence") and 7 ("The Last Judgment") present an interesting contrast. Story 5 exhibits a high proportion of 'Cognition' labels, reflecting a narrative structure that deeply explores the artificial being character's inner world. Whereas Story 7, prominently features 'Judgment'-related vocabulary and thus the label distribution, for it contains substantial content revealing the AI judge's beliefs and decision-making criteria. Both Story 5 and 9 ("Sam-sara") show a predominance of 'Motion' labels, Story 5's greater variance compared to Story 9 suggests differing levels of judgment clarity within this category. It is also noteworthy that 'Communication' labels frequently rank high in prominence, which indicates that many SF works portray artificial beings as capable of linguistic interaction with humans. A detailed breakdown of the total number

of annotations assigned by each individual annotator across the 11 stories is provided in Appendix B. This data enables a comparative analysis of annotators' tendencies in identifying and marking relevant spans, as reflected by the total quantity of labels assigned.

3.4 LLM annotation

The high-context dataset built through human annotation contributes to comparing and improving LLMs' literary comprehension. We created prompts similar to human annotation guidelines, instructing LLMs to mark artificial beings' actions and cognitive processes in the stories and attach appropriate labels.

We utilized Claude 3.5 Sonnet and GPT-4o, both state-of-the-art models known for their excellent multilingual support, including Korean. We conducted zero-shot (providing only guidelines and receiving output in a predetermined format) and few-shot (providing 7 examples) approaches, creating a total of 4 datasets. Our experiments with various lengths³ showed that smaller text units tended to increase the number of labels, but when units became too small, context was lost, resulting in inaccurate character identification and reduced accuracy. We judged that about 5,000 characters was the most appropriate parameter. Stories were divided into 3-6 units, and we integrated the outputted JSON annotation files by work for analysis.

The annotation task involves multiple stages that humans perform intuitively. From a language model's perspective, however, these stages are distinct and sequential:

1. Named Entity Recognition (NER): Distinguishing characters and identifying whether a specific noun/pronoun refers to an artificial being within a story.
2. Verb Span Identification: Accurately identifying the span of Korean verbs that denote actions in the text.
3. Verb Categorization: Categorizing verbs based on contextual meaning.

Comparing and analyzing the annotation data generated through this multi-layered process reveal LLMs' strengths or weaknesses in interpreting complex narrative structures.

To evaluate LLMs' performance of literary com-

³Due to API token limitations, text in segments of about 5,000-6,000 Korean characters were the maximum amount that could be processed at a time without output annotations being cut off.

prehension tasks, we conducted a comprehensive analysis comparing our human-annotated datasets with LLM outputs. Examination of the matching rates between LLM predictions and human annotations was followed by a detailed investigation of the corresponding labels and text segments. Our approach assessed the quality of LLMs' literary analysis within broader contexts, paying particular attention to cases of high agreement and significant discrepancies.

To provide a comprehensive overview of the annotation process, Appendix C includes a table showing the total number of annotations per story by each LLM model, which can be compared with the total number of human annotations.

4 Results and Analysis

4.1 Inter-annotator Agreement Score

Our annotation process consisted of two main steps: 1) precisely marking lexical spans representing artificial beings' actions in the SF texts, and 2) assigning appropriate labels to the marked spans. To assess the reliability of this process, we evaluated inter-annotator agreement using two distinct methods.

The left heatmap in Figure 2 illustrates *sentence-level label agreement* based on Jaccard similarity. This approach matches sentences containing labeled text with the eight labels assigned by each annotator, disregarding detailed text spans. Pairwise agreement ranged from 7.5% to 67.2%, with an overall mean of 42.6%. This wide range coupled with a moderate average indicates significant variability in labeling consistency across annotator pairs. It suggests substantial subjective differences in interpreting artificial beings' actions in literary contexts, while still maintaining a moderate level of overall consensus.

The right graph in Figure 2 depicts *span-based fuzzy label agreement*, a more refined measure. While the average agreement score (53.3%) was higher, it showed greater variability across annotator pairs and works. This analysis scored span agreement as complete match, partial match, or no match, considering both the overlap of annotators' marked text spans and label consistency. Final scores were calculated by verifying label matches for complete and partial span matches, applying appropriate weightings (Equation 1).

Agreement rates calculated by this method ranged widely from 14.0% to 76.2% across an-

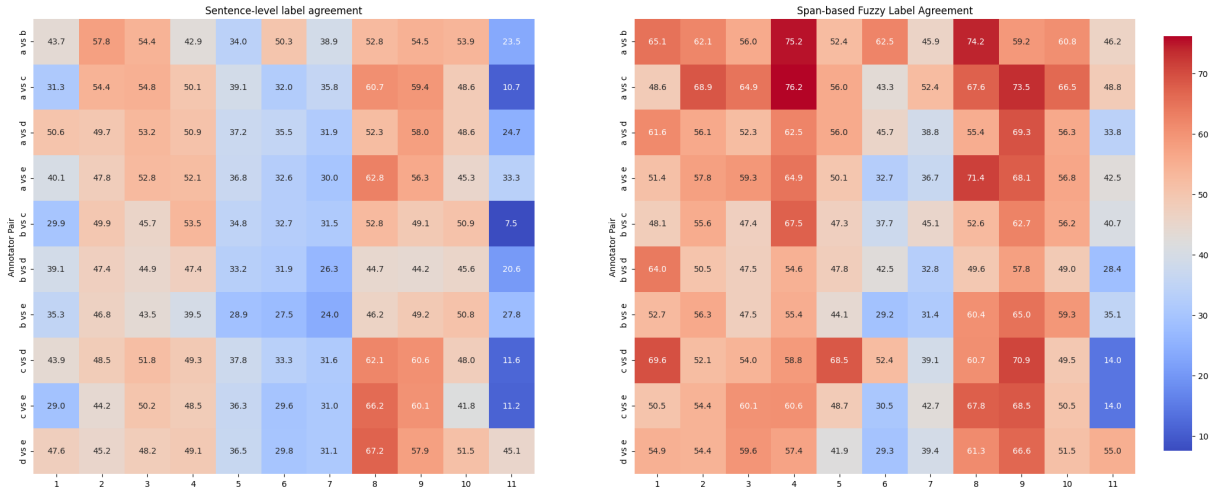


Figure 2: Results of Inter-annotator Agreement Analysis. Left: Heatmap of sentence-level label agreement based on Jaccard similarity. Right: Span-based fuzzy label agreement scores.

notator pairs, reflecting the task’s difficulty and subjectivity. Certain annotator pairs (e.g., a vs. b, a vs. c) consistently showed high agreement⁴, while others (e.g., d vs. e) demonstrated relatively low agreement.

$$\text{Score} = \left(\frac{M_{pe} \times 1 + M_{pa} \times 0.5}{M_{pe} + M_{pa} + M_{non}} \right) \times 100 \quad (1)$$

M_{pe} : number of perfect matches

M_{pa} : number of partial matches

M_{non} : number of non-matches

In most of the texts, the span-based fuzzy method demonstrates generally higher agreement scores compared to the sentence-level method, although it exhibits a wider distribution range. This suggests that the span-based method more effectively captures subtle differences between annotators by accounting for detailed textual elements. The distribution of agreement scores across annotator pairs for each story is provided in the Appendix D.

4.2 LLM Annotation Evaluation

The four types of LLMs labeled 1,000-2,000 data points for each of the 11 stories, similar to the distribution of human annotation data (averaging 1,903 annotations across 5 annotators). To assess how well LLMs understood and classified the actions of artificial beings in the stories without additional training (or to what extent they could match human

⁴The annotator pairs showing high agreement had relevant academic backgrounds: annotator a is the first author of this paper and a master’s student in digital humanities, while annotators b and c hold doctoral degrees in literature.

Model	Total Annotations	Span Matches	Span Unmatches
Claude 3.5 Sonnet (zero shot)	1613	1045 (64.8%)	568 (35.2%)
Claude 3.5 Sonnet (few shot)	1270	878 (69.1%)	392 (30.9%)
GPT-4o (zero shot)	2917	1225 (42%)	1692 (58%)
GPT-4o (few shot)	2775	1305 (47%)	1470 (53%)

Table 3: Summary of Model Annotations and Span Matching.

comprehension), we applied a span-based fuzzy label agreement method similar to the one used in earlier evaluation. We aligned span-marked words and labels sentence by sentence, comparing them with five human annotators’ responses. A Span Match was recorded if there was any overlap (2 or more Korean characters) in the text span. In our analysis, we considered a match to occur when an LLM’s annotation (either span or label) aligned with at least one human annotator. This approach was chosen for both span and label agreement to preserve the diversity of human annotations, which was a key focus of our study. We deliberately avoided creating a “gold standard” based on majority agreement among human annotators, as this would have diluted the individual perspectives we aimed to capture. This method allows us to evaluate LLM performance while acknowledging the inherent variability in human interpretations of textual content. The proportion of entities with matching spans be-

tween LLMs and humans was approximately 60% for Claude 3.5 Sonnet and 40% for GPT-4o. Both models showed fewer mismatched labels in few-shot scenarios compared to zero-shot.

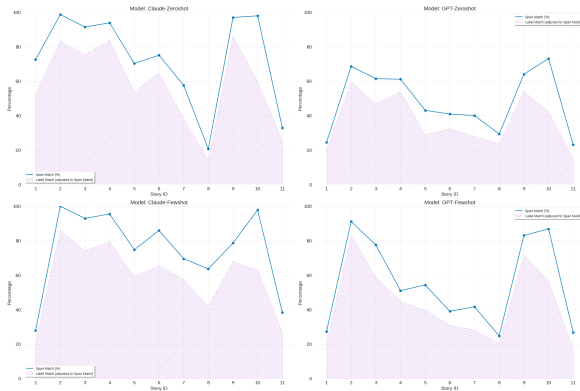


Figure 3: Span Match (line graph) and Label Agreement (shaded area) Rates between 4 different LLMs and Human Annotators across stories, arranged in a clockwise direction starting with Claude-Zeroshot, followed by GPT-Zeroshot, GPT-Fewshot, and ending with Claude-Fewshot.

Figure 3 illustrates the percentage of annotations with matching spans to at least one human reference (line graph) and the proportion of these that also had matching labels with at least one human annotator (shaded area) for each model across stories. Notably, when LLMs correctly identified words describing artificial beings’ actions, the corresponding labels matched human annotators’ labels in over 65% of cases, often exceeding 80%. Specifically, Claude’s model achieved over 75% label agreement in 8 out of 11 stories, while the GPT model reached this threshold in 7 out of 11 stories. This clearly demonstrates that LLMs can detect and classify the actions of artificial beings in stories at a level approaching human annotators.

However, LLMs struggled with Named Entity Recognition (NER), particularly in tracking specific characters. This difficulty likely stems from the varied and often indirect references to artificial entities in the stories’ contexts. The unmatched data from stories 1, 8, and 11, which showed notably low span match rates, mostly consisted of annotations about human characters appearing early in the stories. There were also instances where LLMs misinterpreted passive verbs targeting artificial beings as their actions. These findings suggest that while LLMs excel at sentence-level action classification, they still have room for improvement in consistently tracking characters across broader contexts.

Future research should focus on enhancing contextual understanding and long-term dependency processing to overcome these limitations.

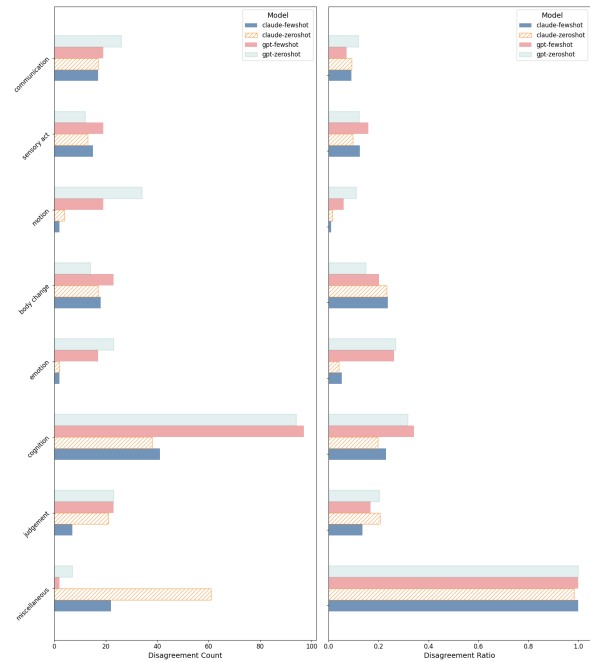


Figure 4: Distribution of label disagreements between LLMs and human annotators for matched spans. The left graph shows the number of label disagreements, while the right graph represents the ratio of label disagreements to the total number of matched spans.

Figure 4 shows the distribution of cases where LLMs identified the matched span but assigned different labels compared to human annotators. The most striking discrepancy occurs in the Cognition category across all models, with GPT models showing an even higher rate of disagreement. Upon closer examination of the text, this divergence is largely centered on machine-related terminology such as ‘record’, ‘upload’, ‘transfer data’, ‘change configuration’, and ‘execute facial recognition’. Human annotators tended to interpret these actions contextually as Motion or Sensory acts, applying a more anthropomorphic perspective to artificial entities. In contrast, LLMs consistently classified these as Cognition, viewing them primarily as computational processes. This discrepancy highlights an intriguing difference in how humans and AI interpret the cognitive processes of artificial beings in literature. Additionally, Claude models, especially Claude Zeroshot, show a notably higher use of the Miscellaneous label, suggesting a more cautious approach to ambiguous actions. Interestingly, GPT models more frequently applied

the Emotion label, while Claude models showed minimal disagreement in this category, indicating varying approaches to emotion recognition between the two model types. These patterns reveal distinct strengths and limitations in how AI models interpret artificial beings' actions, particularly in complex cognitive processes and ambiguous behaviors.

When examining the ratio of label disagreements to the total number of matched span annotations, the overall trends remain similar. However, it's important to note that human annotators were strongly discouraged from using the 'miscellaneous' label (this instruction was also included in the LLM prompts but was not fully adhered to). Consequently, all instances of the 'miscellaneous' label are counted as 'disagreements' in this ratio calculation.

The complete distribution of all labels matching human annotations across different works can be found in Appendix E.

5 Discussion

We investigated the behavioral patterns of artificial beings in contemporary Korean SF short stories. By leveraging both human expertise and large language models (LLMs), we conducted a comprehensive analysis that revealed the complexity and subjectivity involved in categorizing the actions and cognitive processes of artificial characters in literary contexts.

Our novel approach deconstructs the multi-layered cognitive processes inherent in literary comprehension, pioneering a literature-specific framework for evaluating inter-annotator agreement. The higher average agreement score (53.3%) obtained through the span-based fuzzy method, compared to the sentence-level method (42.6%), suggests that considering detailed textual elements better captures nuanced differences and potential consensus in annotators' interpretations. For non-literary text annotation tasks aimed at "predictive accuracy" or "generalizability," an inter-annotator agreement rate around 50% would typically be deemed insufficient. However, given the intricate nature of reading literature, which values diverse interpretations, this agreement rate proves to be a significant finding. The possibility of varied interpretations for the same character demonstrates the depth and richness of literary texts and confirms the active meaning-making processes of readers. The metric we developed presents both unique op-

portunities and methodological challenges for computational approaches in studying representations of characters in fiction.

LLMs demonstrated promising results in matching human annotations, particularly excelling in sentence-level action classification. Claude 3.5 Sonnet and GPT-4o achieved impressive span match rates of approximately 60% and 40% respectively, with label agreement rates frequently exceeding 80% for correctly identified spans. These unprecedented results indicate that LLMs can analyze and categorize artificial beings' actions in literature to nearly human-level accuracy. Notably, few-shot learning approach yielded minimal performance improvements, suggesting that LLMs may already possess specialized capabilities for such high-context tasks, rendering additional 'examples' less critical. However, the models' struggles with Named Entity Recognition (NER) and character tracking across broader contexts highlight areas for improvement in AI's literary comprehension abilities. A particularly significant finding was LLMs' ability to distinguish 'actions' as semantic units rather than merely grammatical verbs (POS). LLMs generally marked necessary Korean objects or auxiliary predicates correctly, enabling clear distinction of artificial beings' behaviors and highlighting the potential in complex literary annotation tasks.

This research provides innovative insights into artificial beings in Korean SF stories, potentially stimulating further studies in this emerging field. Furthermore, our interdisciplinary approach enhances understanding of literary texts, as well as offers valuable insights for developing more sophisticated NLP models capable of grasping contextual nuances and long-term narrative dependencies. These findings present new possibilities for the convergence of literary studies and AI technology.

6 Limitations and Conclusion

This study enhanced annotation robustness by utilizing both human annotators and state-of-the-art LLMs to analyze behavioral patterns of artificial beings in Korean SF works. To ensure methodological transparency, we provided detailed approaches for inter-annotator agreement and LLM performance evaluation. However, identifying and categorizing actions in literary texts remains inherently subjective and is inevitably situated within the contemporary Korean SF contexts and culturally specific understandings of artificial intelligence. In terms

of data, our focus on 11 Korean SF short stories by emerging authors who won a specific literary award potentially restricts the generalizability of our findings. Given the recent proliferation of Korean SF works featuring various artificial beings beyond these stories, expanding the research to include a broader range of contemporary Korean SF literature could have provided more comprehensive insights.

On the technical side, LLMs exhibited limitations potentially stemming from their training data and prompting strategies. The text segmentation necessitated by technical constraints may have affected the models' grasp of overall narrative context. Another limitation is the omission of a detailed linguistic analysis of marked lexical ranges, which was excluded due to space constraints.

Even with these limitations, our findings highlight the complex interplay between literary expressions and technological capabilities, revealing both the potential and limitations of current AI technologies in analyzing nuanced literary contexts. Discrepancies between human annotators and LLMs in interpreting artificial beings' actions underscore the subjective nature of literary analysis and the challenges in AI's comprehension of contextual nuances. However, LLMs's promising performance in sentence-level action classification suggests a path forward for integrating AI tools into literary studies. This research contributes to the ongoing dialogue between science fiction and AI development, offering insights for future studies in both fields.

Acknowledgements

We acknowledge that, due to copyright restrictions on the stories analyzed in this study, we are unable to publicly share the complete dataset. We have endeavored to ensure the utmost transparency in presenting our methodology and results, consistent with these legal constraints.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2024. [Open-source llms for text annotation: A practical guide for model setting and fine-tuning](#). *Preprint*, arXiv:2307.02179.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. [An annotated dataset of coreference in English](#)

[literature](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Sejal Papat, and Sheng Shen. 2019. [An annotated dataset of literary entities](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Parikshit Bansal and Amit Sharma. 2023. [Large language models as annotators: Enhancing generalization of nlp models at minimal cost](#). *Preprint*, arXiv:2306.15766.

Massimiliano Ciaramita and Yasemin Altun. 2006. [Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger](#). In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602, Sydney, Australia. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2023. [Is gpt-3 a good data annotator?](#) *Preprint*, arXiv:2212.10450.

Jens Eder, Fotis Jannidis, and Ralf Schneider. 2010. *Characters in Fictional Worlds*. De Gruyter.

John Frow. 2014. *Character and Person*. Oxford University Press.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. [Annollm: Making large language models to be better crowdsourced annotators](#). *Preprint*, arXiv:2303.16854.

Rebecca M. M. Hicke and David Mimno. 2024. [\[lions: 1\] and \[tigers: 2\] and \[bears: 3\], oh my! literary coreference annotation with llms](#). *Preprint*, arXiv:2401.17922.

Deokgu Hong. 2023. Representations of scientists in contemporary korean science fiction -centered on the novels of kim cho-yeop, shim nae-ul, and jung bo-ra. *Journal of Popular Narrative*, 29(3):69–103.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Jiyong Lee. 2023. Significance and transformation of nonhuman characters in korean science fiction - focusing on science fiction content from the 2010s onwards. *The Society Of Korean Language Culture*, 82:197–225.

Franco Moretti. 2013. *Distant Reading*, volume 93. Verso.

- Haein Oh. 2023. Post-body imagination in south korean sf novels - focused on 「little baby blue pill」 by jung sae-rang and 「laura」 by kim cho-yeop. *Journal of Korean Literary Criticism*, 79:71–104.
- Andrew Piper. 2024. What do characters do? the embodied agency of fictional characters. *Journal of Computational Literary Studies*, 2(1).
- Matthew Sims, Jong Ho Park, and David Bamman. 2019. [Literary event detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.
- Sandeep Soni, Amanpreet Sihra, Elizabeth F. Evans, Matthew Wilkens, and David Bamman. 2023. [Grounding characters and places in narrative texts](#). *Preprint*, arXiv:2305.17561.
- Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. [Improving automatic quotation attribution in literary novels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.
- Aekyung Yoon. 2022. The free will and gender performativity of posthuman in korean sf novel. *International Language and Literature*, 53:81–106.

A

Table 4: List of 11 Science Fiction stories and their details.

Story ID	Title	publication year	Writer	# of syllables	# of sentences	name of the AI in the story	functions
1	피코 Pico	2017	Lee, Gunhyuk	19,089	588	Pico, Freya	Companion AI
2	TRS가 돌보고 있습니다 TRS is Providing Care	2018	Kim, Hyejin	17,961	495	TRS	Care Robot
3	마지막 로그 Last Log	2018	Oh, Jeongyeon	20,168	432	Joy	Euthanasia Assistance Android
4	라디오 장례식 Radio Funeral	2018	Kim, Sunho	16,602	454	Android	Conversation and Service Robot
5	독립의 오단계 Five Stages of Independence	2018	Lee, Ruka	46,906	1,198	I, Model Name A796, Serial Number 04-1963-59	Cyborg Android Integrated with a Human Brain
6	옛날 옛적 판교에 서는 Once Upon a Time in Pangyo	2022	Kim, Kuman	22,461	564	I	In-Game AI
7	최후의 심판 The Last Judgment	2023	Han, Isol	39,208	1,027	Solomon, Solo 3.0	AI Judge
8	두 개의 세계 Two Worlds	2023	Park, Minhyeok	40,461	1,201	Root	Dome Environment Management AI
9	삼사라 Samsara	2023	Jo, Seowol	15,898	316	Sarah, Abe	Artificial Persona of a Spaceship's Main Computer
10	제니의 역 Jenny's Reversal	2023	Choi, Ia	16,403	392	Jenny	Multicultural Family Assistance AI
11	발제자르는 이 배에 올랐다 Balt-hazar Boarded This Ship	2023	Heo, Dallip	18,939	622	Rimey	Privately-Created AI Stored on a Server

B

Table 5: Total number of annotations per story by each human annotator (a-e).

Story ID	a	b	c	d	e	Average
1	146	98	102	127	97	114
2	186	155	206	168	160	175
3	149	158	166	123	141	147
4	121	132	143	123	107	125
5	417	299	366	441	348	374
6	104	87	93	101	49	87
7	302	218	350	228	188	257
8	91	104	100	82	89	93
9	336	318	390	369	366	356
10	88	82	96	91	74	86
11	40	62	204	43	44	79

C

Table 6: Total number of annotations per story by each LLM model.

Story ID	Claude-Zeroshot	Claude-Fewshot	GPT-Zeroshot	GPT-Fewshot	Average
1	80	151	164	103	125
2	72	64	124	101	90
3	81	85	202	138	127
4	80	88	113	98	95
5	352	241	791	499	471
6	68	64	154	249	134
7	276	141	330	543	323
8	170	55	256	267	187
9	159	168	222	253	201
10	47	46	52	53	50
11	228	167	509	421	331

D

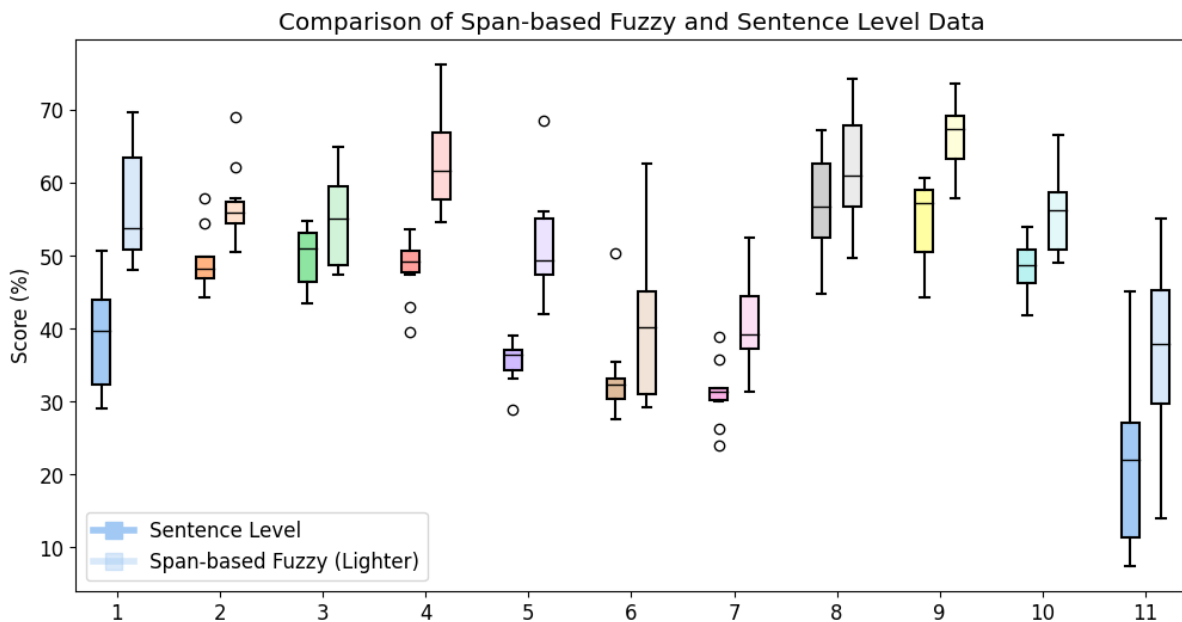


Figure 5: Distribution of Inter-annotator Agreement Scores Across Stories.

E

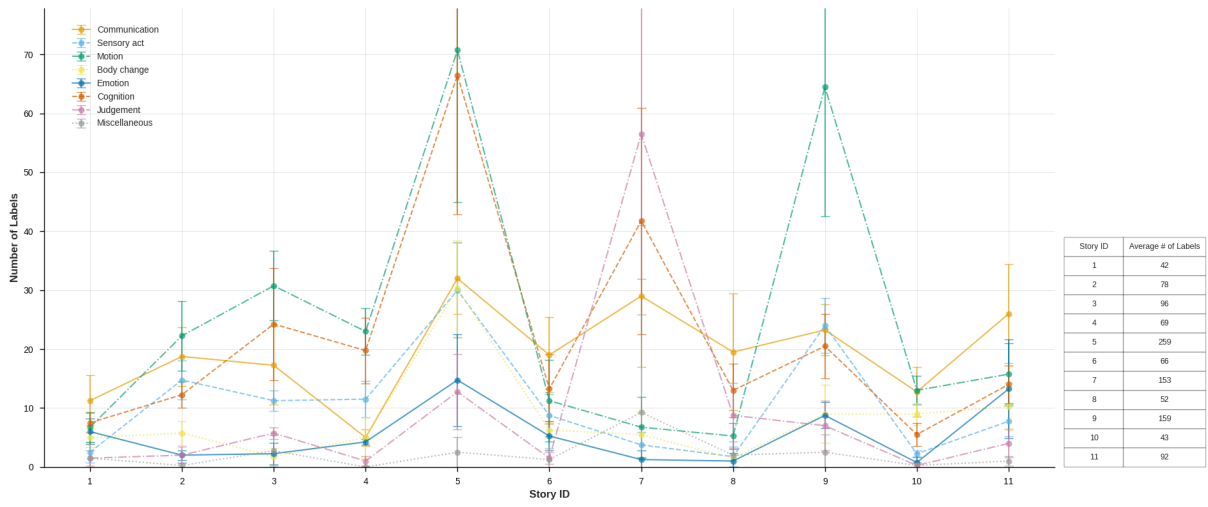


Figure 6: Label distribution of LLM annotations matching human annotation coverage. Values represent the average across four different LLM models.

Text vs. Transcription: A Study of Differences Between the Writing and Speeches of U.S. Presidents

Mina Rajaei Moghadam

Northern Illinois University
mina.rajaei.moghadam@niu.edu

Gülşat Aygen

Northern Illinois University
gaygen@niu.edu

Mosab Rezaei

Northern Illinois University
mosab.rezaei@niu.edu

Reva Freedman

Northern Illinois University
rfreedman@niu.edu

Abstract

Even after many years of research, answering the question of the differences between spoken and written text remains open. This paper aims to study syntactic features that can serve as distinguishing factors. To do so, we focus on the transcribed speeches and written books of United States presidents. We conducted two experiments to analyze high-level syntactic features. In the first experiment, we examine these features while controlling for the effect of sentence length. In the second experiment, we compare the high-level syntactic features with low-level ones. The results indicate that adding high-level syntactic features enhances model performance, particularly in longer sentences. Moreover, the importance of the prepositional phrases in a sentence increases with sentence length. We also find that these longer sentences with more prepositional phrases are more likely to appear in speeches than in written books by U.S. presidents.

1 Introduction

Scholars across various fields have sought to answer what makes writing different from speaking. The answers range from the notion that there is no fundamental difference to the belief that they are entirely distinct domains. These investigations lead to in-depth explorations with different approaches and perspectives, depending on the population or the system under study. For example, some scholars look for answers to support non-native speakers during the language acquisition process, while others attempt to measure the cognitive load through the sound or word production process. Likewise, some try to enhance our ability to program machines and unlock new insight into the differences between spoken and written text.

Rajaei Moghadam et al. (2024) investigated the difference between speaking and writing, focusing on morphological, lexical, and syntactic features at both sentence and chunk levels. They showed the

superiority of BERT (Devlin et al., 2018) as well as the importance of sentence length, percentage of nouns, percentage of verbs, and depth of the parse tree.

In this paper, we expand the corpus and focus on high-level syntactic features to examine their effectiveness in distinguishing the transcriptions of speeches and written books by United States presidents. We analyze linguistically inspired features instead of simply counting categories.

Throughout the paper, the text is analyzed in different sentence lengths, categorized as short, medium, and long. We use CoreNLP (Manning et al., 2014) as a state-of-the-art to parse the sentences. Pinto et al. (2016) provide a comparison of several NLP toolkits, including NLTK, OpenNLP, and Stanford CoreNLP on both formal and informal texts and conclude that depending on the task and text types, the toolkits perform differently.

Regarding text similarity measurement, Wang and Dong (2020) recommend using a combination of techniques and models for higher accuracy, concluding that no single method works best for all similarity measurement tasks in NLP. For this investigation, we use Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), and BERT as our machine-learning classification models to answer the following questions:

RQ1: How do syntactic features impact detection performance in sentences with different lengths?

RQ2: Which syntactic features improve the model performance in distinguishing written sentences from transcribed spoken sentences?

In the literature review, we will explore how other scholars have approached these questions. The methodology will be detailed in the procedure section. Following that, we will present the re-

sults and discussion, draw conclusions, outline the limitations, and finally offer future research. The artifacts of this work are published online.¹

2 Related Work

In this section, we explore related findings in linguistics, cognitive science, and computer science.

2.1 Linguistic and Syntactic Analysis

In comparing writing and speaking, Blankenship (1962) found more general similarity than dissimilarity in sentence patterns with little variation in sentence length, along with mood indicators like imperative. Additionally, Blankenship found more passive constructions in writing. Working with a group of psychology students, Drieman (1962) realized that, compared to speaking, writing has shorter text with more diverse vocabulary and more multisyllabic and longer words. Later, DeVito (1967) analyzed samples of spoken and written language from university professors and concluded that spoken language relies more on verbs and adverbs, while written language uses more nouns and adjectives. These differences were further quantified by analyzing ratios between parts of speech, showing that speech uses fewer qualifiers than written language. O'Donnell (1974) used samples by the same male adult and examined syntactic features like gerunds, passive constructions, and attributive adjectives, which can be found more in written language. Einhorn (1978) also kept the subject and content similar, as she believed this would help understanding the effect of mode of communication. She worked with the writing and recorded speeches of ten famous men, and even though speeches were edited for publication, she still found many differences. For example, they contained more personal references, both singular and plural.

Through multiple attempts to understand why writing and speaking differ, Biber (1986b) identifies three key parameters, interactive vs. edited text, abstract vs. situated content, and reported vs. immediate style, that underlie textual variation in English. Additionally, Biber (1986a) believes that for such studies a comprehensive approach can capture the existing complexity between these two modalities. However, Biber and Grey (2011), in contrast to the conventional view, showed that both conversation and academic writing are grammatically complex, though the sources of complexity

are different. In writing, sentences are compressed due to more use of phrasal expressions, including prepositional phrases as post-modifiers.

2.2 Linguistic and Contextual Influence

Akinnaso (1982) believes these differences are rooted in the objectives of the speaker and writer as well as the communicative and situational context. Chafe (1979), while underscoring the matter of context, identifies integration and involvement as two key distinctions between spoken and written language, meaning that writing is more integrated due to its coherent structure while speaking has a higher involvement rate as speakers are more engaged with the audience. Redeker (1984) believes the four categories of involvement, integration, detachment, and fragmentation work better for such distinctions and similarly refers to speaking as a mode with higher involvement that contains more self-reference items. Poole and Field (1976) highlight that oral language has simpler structures with more adverbial elaboration, which reflects the immediate and personal nature of spoken communication and recalls the importance of communicative context.

2.3 Cognitive Science

Early studies like Woolbert (1922) argue that despite some similarities between these two modes of communication, they are fundamentally different. Woolbert categorizes three processes in writing, thought, language, and typography, while identifying four processes in speaking including thought, language, voice, and action. Therefore, Woolbert counts both production mediums as means to manifest thought. Olson (1996) sees writing as a gateway for studying language and understanding the relationship between writing and cognition.

Liu (2023) reviews the distinction between production, perception, and form. During the production process, voice quality in speaking and, equally, punctuation in writing convey the meaning. Perception deals with the immediacy of feedback. Regarding form, Liu identifies differences in three main areas: lexical richness, grammar, and structure. Trying to explore the possibility of language measurement, Fairbanks (1944) and Mann (1944) worked on two groups: freshman students and individuals with schizophrenia. Both employ methods such as type-token ratios and grammatical analysis, including examining prepositions and conjunctions. The most notable difference, as Fairbanks men-

¹<https://github.com/mosabrezaei/Text-vs.-Transcription>

tions, is the increased use of personal pronouns by patients with schizophrenia.

In another study, [Rezai \(2022\)](#) works on productions of individuals with primary progressive aphasia and finds that familiarity with terms and topics decreases the cognitive load, thus easing the production process by either use of complex syntax structure with simple vocabulary or the opposite. [Cleland and Pickering \(2006\)](#) notice the use of syntactic priming, meaning that the speaker tends to reuse previously used syntactic structure. Their results show that during the production process, syntax is accessed the same way in both speaking and writing, suggesting a similar underlying cognitive mechanism.

[DeVito \(1966\)](#) highlights that writing has greater verbal diversity than speaking due to differences in the encoding process, including time constraints and pressure on the speaker when uttering a sentence, which is the reason speakers use more familiar and shorter words. Such differences show the distinct cognitive and linguistic demands on individuals through the production process. Likewise, [Chafe and Tannen \(1987\)](#) look at structural differences, cognitive implications, and social functions. They similarly refer to the immediate and context-dependent characteristics of speaking as opposed to writing. [Gray and Biber \(2013\)](#) analyze lexical frames in academic prose and conversational English. Their study shows that writing tends to use more grammatical structure and function word-based frames, while in conversation, fixed, verb-based frames are more common, which also, like DeVito, reflects the immediate and interactive aspects of spoken communication.

2.4 Computer Science

While [Biber \(2020\)](#) invites more investment in phonetic and phonological corpora to help in studies of speaking, [Pangtay-Chang \(2009\)](#) shows that text-based computer-mediated communication is becoming similar to what we produce in oral communication.

Understanding differences in writing and speaking will serve other areas of research like human-human-computer interactions. With this focus, [Akhtiamov et al. \(2017\)](#) analyzed speech through acoustical, syntactical, and lexical lenses. Ultimately, their study suggests a greater reliance on conversational context rather than acoustic cues. Similarly, [Balagopalan et al. \(2020\)](#) use NLP to de-

tect symptoms of Alzheimer's disease (AD), which impacts both the content and acoustics of spontaneous speech. This study reveals that fine-tuned BERT models outperform traditional feature-based methods in detecting cognitive impairments associated with AD.

Such exploration will also be useful for related research in stylistics. [Blankenship \(1962\)](#) concluded that the formation of a syntactic structure is a matter of individual style; therefore the medium of delivery, whether writing or speaking, has minimal influence. [Kurzynski \(2023\)](#), through an analysis of perplexity, systematicity, and characteristic words of Mao Zedong, introduces these metrics as helpful ones to understand Mao's writing style. In another study, [Freedman \(2017\)](#) employs syntactic and bag-of-words approaches to distinguish different sections of the book of Isaiah. Also, [Freedman and Kriegbaum \(2014\)](#) used features like prepositional phrases along with machine learning techniques to investigate student responses. Expanding on stylistics, [Khalid and Srinivasan \(2020\)](#) used 262 stylistic features to analyze style across nine online communities to explore the importance of style in these communities rather than individual style. They found higher accuracy in style-based prediction as opposed to content-based predictions of community membership, particularly in smaller data sets.

[Rajaei Moghadam et al. \(2024\)](#) study syntactic and non-syntactic features to identify the most important ones for detecting spoken and written textual data. However, their study did not examine high-level constructs like prepositional phrases. [Katre \(2019\)](#), with a discourse analysis approach, used NLTK and Matplotlib to process a large corpus of political speeches to create visual tools like lexical dispersion plots, time-series plots, word clouds, and bar graphs.

[Berriche and Larabi-Marie-Sainte \(2024\)](#) examine writing style differences between human and ChatGPT-generated content. They employed classical classifiers and ensemble methods, training them with over 30 stylometric features. They extracted lexical and syntactic features including the frequency of conjunctions, pronouns, and prepositions. Through multiple experiments, they concluded that the ensemble learning classifiers outperformed the classical classifiers. Regarding style generation, [Montfort et al. \(2021\)](#) focus on generating narrative style (not the plot) with referring ex-

pressions. In other words, they explore how changing the referring expressions can model different literary styles. By keeping all other influential elements in the discourse constant and changing only reference conventions, they emphasize the use of nouns and noun phrases for generating different writing or narration styles.

3 Procedure

In this section, we describe the dataset and the extracted features.

3.1 Dataset

As outlined in the future work section of [Rajaei Moghadam et al. \(2024\)](#), we aimed to extend the number of extracted sentences. Therefore, in this study, we have expanded the corpus volume, which now contains 41,306 sentences, comprising 20,654 spoken samples and 20,652 written samples, compared to the earlier dataset of 13,600 spoken and 13,600 written samples.

We obtained transcriptions of spoken language from [Miller Center of Public Affairs University of Virginia \(2022\)](#), which covers transcriptions from George Washington to the present time. For the writing samples, we used ten complete books written by presidents, three of which we obtained from [Project Gutenberg \(n.d.\)](#).

To ensure the accuracy of calculations, all the pages that were not part of the main content were removed. Furthermore, multiple whitespaces were changed into single whitespaces. For sentence extraction, we used the *nlk* library ([Bird et al., 2009](#)), while CoreNLP (version 4.5.7) was employed for tokenization and word counting.

3.2 New Features

In the exploration of what exactly makes writing and speaking different, there is no single definite answer. Therefore, in addition to utilizing some of the features from [Rajaei Moghadam et al. \(2024\)](#), we will examine the following six features:

- Pronoun and noun phrases in the subject
- Passive and active sentences
- Comparative and superlative
- Imperative structures
- Conjunction phrases
- Prepositional phrases

3.2.1 Pronoun and Noun Phrase in Subject

We examined syntactic subjects to determine whether they were occupied by noun phrases (NP) or pronouns (PRN). [Rajaei Moghadam et al. \(2024\)](#) counted noun phrases and personal pronouns as separate features. In this paper, we only consider these two elements in the subject position. Such analysis deepens our understanding of nominal construction and sentence complexity in both modalities.

According to [de Marneffe and Manning \(2008\)](#), a nominal subject (*nsubj*) refers to a noun phrase that is the syntactic subject of a clause. Here, we use a combination of the parse tree and the Enhanced Dependency subsystem of Stanford CoreNLP to identify nominal subjects and their referents with higher accuracy.

3.2.2 Passive and Active Sentences

According to [Aygen \(2016\)](#), the active voice is the typical form in which the subject of the sentence is the agent. To do this, PassivePy package ([Sepehri et al., 2023](#)) in the SpaCy library ([Honnibal et al., 2020](#)) enables us to compute active, agentless passive, and agentive passive forms.

3.2.3 Comparative and Superlative

The comparative form is used to compare two sets of entities, whereas the superlative form compares more than two sets of entities or groups ([Aygen, 2016](#)). We extracted comparative and superlative structures with JJR, JJS, RBR, and RBS tags from the dataset using Stanford CoreNLP. This extraction includes irregular forms, such as "good", "well", and "best", in addition to those that end with "-er" and "-est" or contain indicators like "more" and "most".

3.2.4 Imperative

The imperative mood is used in direct requests or commands. According to [Aygen \(2016\)](#), imperatives do not have tense or aspect markers and have an implied subject (you). Therefore, this analysis focuses on structures without a stated subject and verbs without tense or aspect modifiers such as gerunds. To achieve this, we use StanfordCoreNLP to extract only sentences that begin with a VB tag.

It is common to find fragments and informal questions in spoken language that start with the base form of the verb, such as "Want fries?", which could be counted as an imperative structure. To address this, we examined the role of punctuation and ultimately decided to only consider sentences

that start with a verb and end with a period or exclamation mark.

In other cases, phrases like "Sleep well, gentlemen" may structurally appear as imperatives but are not interpreted as commands. Similarly, some proverbs and idiomatic expressions use the base form of the verb without an explicit subject, much like imperatives, e.g., "Hit the nail on the head". Although these cases offer interesting avenues for further analysis, they fall outside the scope of this research since our primary focus is syntactic analysis.

3.2.5 Conjunction Phrases

Conjunctions are connector words that link two words, phrases, clauses, or sentences (Zokirjon kizi, 2023). While Rajaei Moghadam et al. (2024) focused on the percentage of coordinators, this study investigates conjunction phrases with the assistance of the Stanford CoreNLP parse tree. The work by de Marneffe and Manning (2008) notes that the parser does not account for symmetrical relations, meaning that we do not observe two conjunction phrase tags (CONJP) in cases like correlative conjunctions, e.g., "not only...but also...". Based on this, we check only for the presence of a single CONJP tag in the parse tree.

3.2.6 Prepositional Phrases

Among the different parts of speech, prepositions are considered as function words and prepositional phrases as grammatical units that act as connectors, typically with noun phrases. They precede or follow other phrases or elements in a sentence to create another phrase or constituent. According to Benelhadj (2015), prepositional phrases exhibit varying levels of structural complexity and generally cannot be understood without considering other elements of the sentence.

For this feature, we first extract and calculate the percentage of each sentence occupied by prepositional phrases (PP) using the Stanford CoreNLP tags. Then, we calculate the percentages of words with PP tags that modify verbs vs. nouns or other parts of speech. In this calculation, words in nested PPs are labeled according to their closest parent.

3.3 Pre-existing Features

In this paper we continue to use the features in Rajaei Moghadam et al. (2024). The features are listed below by category.

Morphological aspects:

- Average syllables per word
- Average words per sentence
- Average characters per word

Lexical aspects of sentences:

- Number of words in a sentence
- Percentage of POS
- Percentage of personal pronouns

Syntactical aspects:

- Percentage of subordinate clauses
- Depth of parse tree
- Percentage of noun phrases
- Average length of noun phrases
- Yes/no questions
- Direct wh-questions

4 Experiments

In this section, the two experiments that we conducted will be described. The first focuses on evaluating syntactic features in sentences with different lengths, and the second analyzes both low- and high-level syntactic features.

4.1 First Experiment

In the first experiment, we evaluated the impact of sentence length on model performance. Given the important role of length (Rajaei Moghadam et al., 2024), we used this insight to minimize model dependency on sentence length. The goal was to determine whether the models performed better when trained on the entire dataset or when focused on specific sentence lengths.

We divided our dataset into three categories based on sentence length: sentences with 18 or fewer words were classified as short, those with more than 18 and up to 37 words as medium, and those with more than 37 words as long. The boundary numbers that define short, medium, and long sentences were determined based on the data distribution to ensure a sufficient number of samples in each category. Then we trained each model on each section, utilizing both syntactic and non-syntactic features.

4.2 Second Experiment

In the second experiment, we shift our focus from sentence length to features. In this experiment, we evaluate the effectiveness of the combination of features in sections 3.2 and 3.3. We ran four

models, SVM, DT, RF, and BERT, on the new feature set. The comparison of these results with the results from [Rajaei Moghadam et al. \(2024\)](#) will determine whether combining features can improve model accuracy. It is important to note that in both experiments the BERT model examines words in sequence but does not utilize any of these features, while the other models are used to analyze the features and their role in distinguishing spoken from written language.

5 Results and Discussion

The results of our first experiment, shown in Figure 1, indicate that accuracy across all models is lower for short sentences compared to longer ones. This suggests two key points.

First, the extracted syntactic features are more informative in longer sentences, and their rare occurrence in short sentences leads to lower performance across all models. Second, despite having more short samples in the dataset, the selected features performed better on longer sentences. This implies that there is likely to be higher accuracy in a larger dataset containing more long sentences.

It should also be noted that the similarity between the results for short sentences and the overall dataset is due to the large number of short sentences, which biases the models' performance. Furthermore, the greater similarity between the results for medium and short sentences, compared to medium and long sentences, is due to the closer boundary numbers for short and medium sentences. Although BERT does not have explicit access to the linguistic features, we note that it performs better than any of the other models.

As shown in Table 1, although the length is the most important feature, its importance decreases as the sentence length increases. On the other hand, as sentences get longer the importance of prepositional phrases (PP, PP_NP, PP_VP) increases significantly.

Looking at RQ2, Table 2 and Figure 2 complement each other. We compare accuracy metrics for spoken vs. written sentences in each model. Table 2 shows that models trained with high-level syntactic features alone tend to have slightly lower accuracy, partially because length is not included in the high-level features and partially because the high-level feature set contains fewer features. Moreover, as shown in the last column, combining all features improves the performance of the models, with Ran-

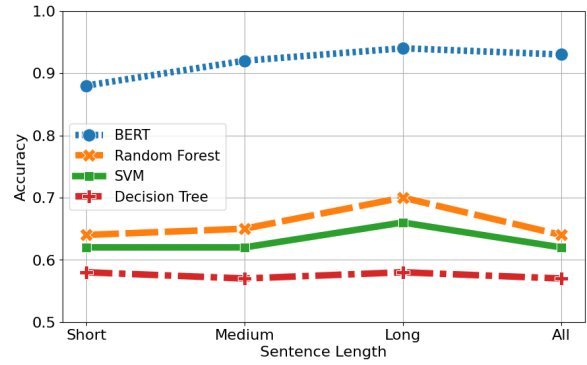


Figure 1: The performance of the models on sentences with different lengths.

dom Forest outperforming the others.

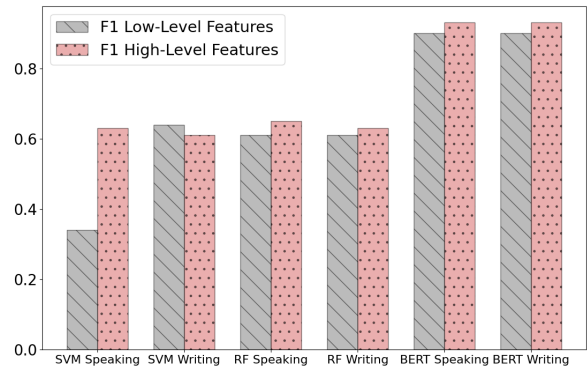


Figure 2: Comparing the performance of models trained with low-level and high-level features.

Figure 2 compares the F1 scores of the SVM, RF, and BERT models for low-level and high-level features. The figure shows improvement in almost all performance metrics. Notably, the significant improvement in the SVM performance for the speaking class is particularly striking. This improvement may be attributed to several factors, such as an increased sample size, the use of different versions of CoreNLP, better feature normalization, and improved feature extraction methods. On the other hand, there is a slight reduction in performance for the writing class in the SVM model, which could indicate that the models are becoming more stable and less biased.

Table 3 presents a comparison between trained models with all syntactic features as well as the BERT model, which corresponds to the "All Features" columns in Table 2. When comparing Table 3 with a similar table in [Rajaei Moghadam et al. \(2024\)](#), we observe an overall improvement in model performance.

Notably, as sentences become longer, the perfor-

Table 1: The report of the six most important extracted features across four different sentence lengths. The numbers indicate the percentage of importance for each feature.

	Rank	Short		Medium		Long		All	
DT	1	Length	0.158	Length	0.115	PP	0.089	Length	0.117
	2	Verb	0.096	Noun	0.075	Length	0.086	Noun	0.084
	3	Noun	0.096	PP	0.070	PP_VP	0.073	Verb	0.077
	4	Words	0.075	Subord	0.069	PP_NP	0.070	Words	0.067
	5	Adverb	0.058	Verb	0.068	Words	0.066	PP	0.063
	6	PP	0.054	PP_NP	0.064	Verb	0.065	D_Tree	0.061
RF	1	Length	0.122	Length	0.095	PP	0.093	Length	0.094
	2	Verb	0.097	Noun	0.072	Length	0.087	Noun	0.075
	3	Noun	0.084	PP	0.070	PP_NP	0.074	Verb	0.075
	4	Words	0.065	PP_NP	0.068	PP_VP	0.068	Words	0.069
	5	D_Tree	0.059	Verb	0.065	Noun	0.066	PP	0.064
	6	PP	0.052	Subord	0.059	Subord	0.066	PP_NP	0.061

Table 2: The performance of the models with different levels of features. The low-level feature data comes from Rajaei Moghadam et al. (2024).

		Low-level Features		High-level Features		All Features	
		Precision	Recall	Precision	Recall	Precision	Recall
SVM	Spoken	0.59	0.67	0.57	0.57	0.61	0.65
	Written	0.63	0.54	0.58	0.58	0.64	0.59
DT	Spoken	0.56	0.56	0.55	0.49	0.56	0.57
	Written	0.57	0.58	0.55	0.60	0.58	0.57
RF	Spoken	0.62	0.69	0.58	0.52	0.63	0.67
	Written	0.66	0.58	0.58	0.64	0.66	0.62

mance of our syntactic features improves, probably because longer sentences provide more information, which enables the models to more accurately distinguish between speech and writing.

Figure 3 shows the feature importance ranking of the merged set of features. Note that high-level features like conjunction phrases (CONJP) and imperatives show less influence. This is possibly due to the fact that these features rarely appear in sentences in our dataset. For instance, the CONJP feature appears in only about 2 percent of all sentences.

As shown in Table 1 and Figure 3, two features stand out in distinguishing written text from the transcribed spoken text: length for all sentences and the percentage of PP for long sentences. The results show that longer sentences and a higher percentage of prepositional phrases appear more frequently in speech than in written books. In other words, U.S. presidents tend to use longer sentences and more prepositional phrases in their speeches than in their books. We conducted a statistical anal-

ysis and visualized the distribution of each of these features to better understand the relationship between these features and the classes of written and spoken texts. As expected, the t-tests in Table 4 show large absolute values and extremely small p-values for both features, indicating significant differences between the two classes. On the other hand, the small negative correlation values in Table 4 and the slight differences in class distribution as shown in Figure 4 indicate that increasing the percentages of these features decreases the probability of labeling a sentence as writing.

6 Conclusions

In this study, we analyze low-level and high-level syntactic features to identify the differences between the speeches and written books of presidents of the United States. We conducted two experiments to achieve these goals.

In the first experiment, sentences were divided into three categories: short, medium, and long. We found that, despite having fewer samples for long

Table 3: Comparing the results of trained models with all syntactic features and the BERT model.

	Labels	Precision	Recall	F1
SVM	Spoken	61%	65%	63%
	Written	64%	59%	61%
DT	Spoken	56%	58%	57%
	Written	58%	56%	57%
RF	Spoken	63%	67%	65%
	Written	66%	62%	63%
BERT	Spoken	92%	93%	93%
	Written	93%	92%	93%

Table 4: Statistical test on the length for all sentences and PP for long sentences.

	Length (all)	PP (long)
t-statistic	12.00	16.43
p-value	3.979×10^{-33}	9.029×10^{-60}
correlation	-0.05	-0.17

sentences, accuracy improves across all models. Increasing the sentence length also raised the importance ranking of prepositional phrases. Furthermore, the most significant features identified are sentence length, verb percentages, noun percentages, and prepositional phrases.

In the second experiment, we added a new set of syntactic features to morphological, lexical, and other syntactic features. The results showed that combining both groups of features improves model performance. Furthermore, sentence length and prepositional phrases emerged as the two important features in distinguishing the textual data of U.S. presidents. Based on our analysis, U.S. presidents are more likely to use prepositional phrases and longer sentences in their speeches than in their books.

7 Limitations

Although the dataset is balanced, we encountered some imbalanced features that appear rarely in sentences. For instance, there were only 351 imperative sentences, which account for less than 1 percent of all sentences. This limitation could affect future work in identifying effective features for this task.

Another limitation is the number of long sentences. By increasing the number of long sentences, or balancing with that of short sentences, we might

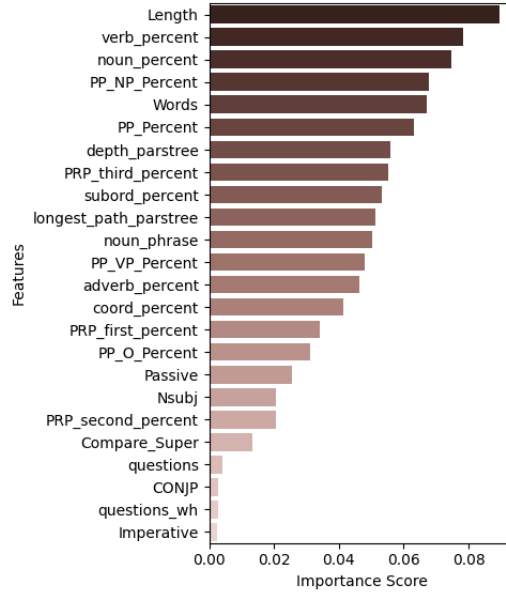


Figure 3: The importance of features in Random Forest model.

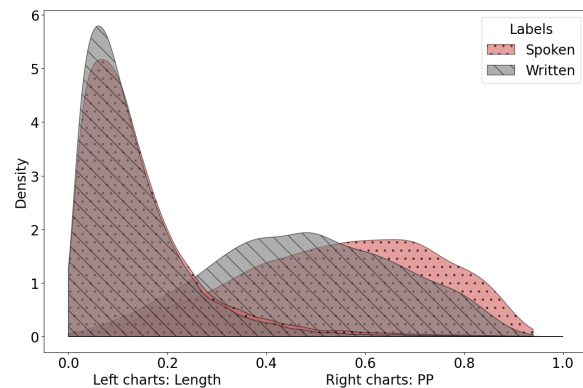


Figure 4: Distribution of the length and PP for each class using kernel density estimation (KDE).

observe higher model performance and allow for the extraction of more accurate patterns.

8 Future Work

Explaining the differences between transcribed spoken and written text is an open area of research, with each study revealing more possible directions for future work. For example, our study demonstrates the weak performance of models on short sentences. For future work, new features need to be introduced and extracted in order to improve the model performance on short sentences.

We are going to perform a deeper analysis of the relationship between sentence length and its impact on both low- and high-level features. We will shift from a categorical approach to a regression-based analysis of sentence length. This means that instead

of categorizing sentences into three groups (short, medium, long), we will analyze the effects across the full range of sentence lengths.

Based on the importance of prepositional phrases, we plan to expand our analysis and study nested prepositional phrases. Additionally, we aim to apply deeper analysis to the different types of prepositional phrases introduced in this study, such as those modifying verb phrases or noun phrases.

References

- Oleg Akhtiamov, Maxim Sidorov, Alexey A Karpov, and Wolfgang Minker. 2017. Speech and text analysis for multimodal addressee detection in human-human-computer interaction. In *Interspeech*, pages 2521–2525.
- F Niyi Akinnaso. 1982. On the differences between spoken and written language. *Language and Speech*, 25(2):97–125.
- Gulsat Aygen. 2016. *English Grammar: A Descriptive Linguistic Approach*, third edition. Kendall Hunt.
- Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer’s disease detection. *arXiv preprint arXiv:2008.01551*.
- Fatma Benelhadj. 2015. Prepositional phrases across disciplines and research genres: A syntactic and semantic approach. Doctoral dissertation, Department of English, University of SFAX.
- Lamia Berriche and Souad Larabi-Marie-Sainte. 2024. *Unveiling ChatGPT text using writing style*. *Heliyon*, 10(12):e32976.
- Douglas Biber. 1986a. On the investigation of spoken/written differences 1. *Studia Linguistica*, 40(1):1–21.
- Douglas Biber. 1986b. *Spoken and written textual dimensions in English: Resolving the contradictory findings*. *Language*, 62(2):384–414.
- Douglas Biber. 2020. Corpus analysis of spoken discourse. *Pronunciation in Second Language Learning and Teaching Proceedings*, 11(1).
- Douglas Biber and Bethany Grey. 2011. Is conversation more grammatically complex than academic writing? In *Grammatik und Korpora 2009: Dritte Internationale Konferenz*, pages 47–61.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly.
- Jane Blankenship. 1962. *A linguistic analysis of oral and written style*. *Quarterly Journal of Speech*, 48(4):419–422.
- Wallace Chafe. 1979. Integration and involvement in spoken and written language. In *2nd Congress of the International Association for Semiotic Studies*, pages 195–215.
- Wallace Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology*, 16(1):383–407.
- Alexandra A Cleland and Martin J Pickering. 2006. Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language*, 54(2):185–198.
- Marie-Catherine de Marneffe and Christopher D. Manning. 2008. *Stanford Typed Dependencies Manual*. Revised for the Stanford Parser v. 3.7.0 in September 2016.
- Joseph A DeVito. 1966. The encoding of speech and writing. *Communication Education*, 15(1):55–60.
- Joseph A DeVito. 1967. *A linguistic analysis of spoken and written language*. *Communication Studies*, 18(1):81–85.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Gerard HJ Drieman. 1962. Differences between written and spoken language: An exploratory study. *Acta Psychologica*, 20:36–57.
- Lois Einhorn. 1978. Oral and written style: An examination of differences. *Southern Journal of Communication*, 43(3):302–311.
- Helen Fairbanks. 1944. *II. The quantitative differentiation of samples of spoken language*. *Psychological Monographs*, 56(2):17–38.
- Reva Freedman. 2017. Can natural language processing help identify the author(s) of the book of Isaiah? In *30th International FLAIRS Conference*, pages 297–300.
- Reva Freedman and Douglas Kriegbaum. 2014. Effects of rewriting essays on linguistic measures of complexity. In *25th Annual Meeting of the Society for Text and Discourse*.
- Bethany Gray and Douglas Biber. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1):109–136.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. doi: 10.5281/zenodo.1212303.
- Paritosh D Katre. 2019. NLP based text analytics and visualization of political speeches. *International Journal of Recent Technology and Engineering*, 8(3):8574–8579.

- Osama Khalid and Padmini Srinivasan. 2020. [Style matters! Investigating linguistic style in online communities](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):360–369.
- Maciej Kurzynski. 2023. The stylometry of Maoism: Quantifying the language of Mao Zedong. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 76–81.
- Yijun Liu. 2023. [Differences between spoken and written English](#). *Communications in Humanities Research*, 3:757–761.
- Mary Bachman Mann. 1944. III. The quantitative differentiation of samples of written language. *Psychological Monographs*, 56(2):39–74.
- Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- Miller Center of Public Affairs University of Virginia. 2022. Presidential speeches: Downloadable data. Accessed: 2022-03-17, Available at <https://data.millercenter.org>.
- Nick Montfort, Ardalan SadeghiKivi, Joanne Yuan, and Alan Y Zhu. 2021. Using referring expression generation to model literary style. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 64–74.
- Roy C O’Donnell. 1974. Syntactic differences between speech and writing. *American Speech*, 49(1/2):102–110.
- David R Olson. 1996. Towards a psychology of literacy: On the relations between speech and writing. *Cognition*, 60(1):83–104.
- Yolanda Pangtay-Chang. 2009. IM conversations in Spanish: Written or oral discourse? *Illinois Language and Linguistics Society 1 (ILLS)*.
- Alexandre Pinto, Hugo Gonçalo Oliveira, and Ana Oliveira Alves. 2016. Comparing the performance of different NLP toolkits in formal and social media text. In *5th Symposium on Languages, Applications and Technologies (SLATE’ 16)(2016)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.
- Millicent E Poole and TW Field. 1976. A comparison of oral and written code elaboration. *Language and Speech*, 19(4):305–312.
- Project Gutenberg. n.d. Project Gutenberg. Retrieved February 21, 2016, from <https://www.gutenberg.org>.
- Mina Rajaei Moghadam, Mosab Rezaei, Miguel Williams, Gülşat Aygen, and Reva Freedman. 2024. [Investigating lexical and syntactic differences in written and spoken English corpora](#). *Proceedings of the 37th International FLAIRS Conference*.
- Gisela Redeker. 1984. On differences between spoken and written language. *Discourse Processes*, 7(1):43–55.
- Neguine Rezaii. 2022. The syntax-lexicon tradeoff in writing. *arXiv preprint arXiv:2206.12485*.
- Amir Sepehri, Mitra Sadat Mirshafiee, and David M Markowitz. 2023. PassivePy: A tool to automatically identify passive voice in big text data. *Journal of Consumer Psychology*, 33(4):714–727.
- Jiapeng Wang and Yihong Dong. 2020. Measurement of text similarity: A survey. *Information*, 11(9):421.
- Charles H Woolbert. 1922. Speaking and writing—A study of differences. *Quarterly Journal of Speech*, 8(3):271–285.
- Tukhtasinova Zarina Zokirjon kizi. 2023. [Conjunctions in English](#). *Modern Science and Research*, 2(9):29–35.

Mitigating Biases to Embrace Diversity: A Comprehensive Annotation Benchmark for Toxic Language

Xinmeng Hou

Columbia University

fh2450@tc.columbia.edu

Abstract

This study introduces a prescriptive annotation benchmark grounded in humanities research to ensure consistent, unbiased labeling of offensive language, particularly for casual and non-mainstream language uses. We contribute two newly annotated datasets that achieve higher inter-annotator agreement between human and language model (LLM) annotations compared to original datasets based on descriptive instructions. Our experiments show that LLMs can serve as effective alternatives when professional annotators are unavailable. Moreover, smaller models fine-tuned on multi-source LLM-annotated data outperform models trained on larger, single-source human-annotated datasets. These findings highlight the value of structured guidelines in reducing subjective variability, maintaining performance with limited data, and embracing language diversity.

Content Warning: This article only analyzes offensive language for academic purposes. Discretion is advised.

1 Introduction

To properly offer people the option to avoid potentially offensive language while also protecting minoritized language varieties from being misidentified, accurate detection that can identify languages despite changes over time is required. Current datasets typically employ multifaceted methodologies for content categorization, taking into account not just the presence of offensive language but also its context, target, and underlying intent (Zampieri et al., 2019; Basile et al., 2019; Mollas et al., 2020). Abusive, toxic, or offensive language and hate speech were often directly identified based on finite lists of phrases (Davidson et al., 2017), annotators' interpretation of the textual content (de Gibert et al., 2018; Founta et al., 2018; Sap et al., 2019; Susanto et al., 2024), or a combination of both (Vargas

et al., 2021; Basile et al., 2019). This raises the issue of an unclear research subject characterized by inconsistencies in terminology and categorization (Fortuna et al., 2020). For instance, hate speech is often treated as equivalent to offensive or toxic language (Susanto et al., 2024), which leads to problems where language that is less offensive than hate speech may be incorrectly classified as non-offensive.

Biases in annotation refer to the systematic tendency of human annotators that leads to errors or skewed labels in the training data used for machine learning models (Davani et al., 2023). The most common approach for mitigating annotator bias is diversifying annotation teams and increasing annotation on each raw piece (Davani et al., 2023; Sap et al., 2019; Geva et al., 2019). However, no research addresses how diverse the annotator team should be and how many annotators were required to eliminate bias efficiently. While diversification and scale help address bias, the root issue often lies in subtle differences in interpretations addressing complex socio-cultural dynamics that are especially vulnerable (Kuwatly et al., 2020). Therefore, rather than treating annotator disagreement as mere "noise" or using majority vote labels to cover up disagreement, inevitable disagreements should be adequately addressed in annotation (Davani et al., 2023, 2021). The main research question is **how to reveal the underlying patterns while minimizing the impact of biased annotations against non-standard language use during the data labeling process to protect language diversity**. Moreover, data may be limited or nonexistent, particularly for endangered dialects, minority language use (Liu et al., 2022), and low-resource scenarios. The second question explores **whether annotated features can improve models' robustness against small datasets and varied language use, making them more accommodating of English variety**. Finally, we observed that skilled and well-trained

human annotators are not always readily available. Instead of relying on untrained annotators who lack expertise in language or social studies, we investigate **whether prompted large language models (LLMs) can serve as a viable alternative.**

As shown in Figure 1, our research addresses three key components: (1) proposing criteria for a prescriptive annotation framework that will be introduced in methodology, (2) conducting a small-scale statistical analysis to compare the framework with the descriptive paradigm and evaluate the performance of prescriptively-prompted LLMs, and (3) testing the framework under limited conditions, using smaller datasets with complex language features without human annotators.

To assess annotation quality, we compared inter-rater reliability across three sets: 400 pieces from the Davidson et al., 2017 dataset following general definitions, our descriptive annotations simulating Davidson et al., 2017 annotations, and our prescriptive annotations on the same 400 pieces. LLMs, prompted based on the prescriptive framework, were used in place of professional annotators to simulate limited human resources. The experiments demonstrate the effectiveness of smaller models fine-tuned on LLM-based prescriptive annotations for a 1942-piece set, comparing their performance to models fine-tuned on unused Davidson et al., 2017 annotations. Key contributions and findings are outlined below:

1. This research proposes a prescriptive annotation benchmark to enable consistent offensive language data labeling with high reliability while preventing biases against language minorities, hence protecting natural language diversity.
2. This research contributes two newly annotated offensive language detection datasets created based on the proposed annotation benchmark ¹.
3. The proposed criteria lead to a higher inter-annotator agreement and reliability between prescriptive human annotations and between prescriptive human annotations and annotation generated by LLMs with prescriptive prompts derived from the annotation benchmark, compared to the original annotations based on vague and descriptive annotation instructions.
4. Smaller models fine-tuned on a multi-source dataset annotated by LLMs outperform models

¹Paparare/toxic_benchmark_2024

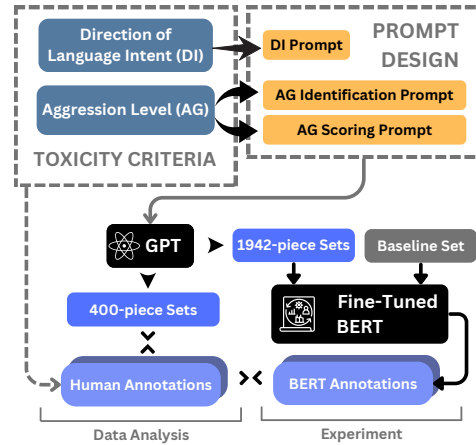


Figure 1: **Research Design:** This research establishes standardized criteria for toxic language annotation and analyzes inter-annotator reliability. Experiments on BERT models across language types tend to demonstrate the broader applicability of the proposed annotation criteria, even with limited resources.

trained on a single, significantly larger dataset annotated by humans, showing the effectiveness of structured guidelines in maintaining performance with limited data size and heterogeneous language types.

2 Related Works

2.1 Common Annotation Bias in Past Datasets

The issue of non-offensive language being mislabeled as offensive is also called unintended bias (Dixon et al., 2018) or, more specifically, lexical bias (Garg et al., 2023) or linguistic bias (Fan et al., 2019) (Tan and Celis, 2019). For example, both (1) and (2) were identified as offensive:

(1) And apparently I'm committed to going to a new level since I used the key. Well FUCK. Curiosity killed the Cat(hy) (Barbieri et al., 2020)

(2) I ain't never seen a bitch so obsessed with they nigga😂" I'm obsessed with mine 😑 (Davidson et al., 2017)

In (1), F**K is used as emotional emphasis. Similarly, slang does not always induce toxicity, as presented in (2); race-related term n***a is a neutral word often found in African American English (AAE) and gender-related b***h. The appropriateness of these terms varies, and their potential to harm others depends on their perlocutionary effect,

influenced by the context and circumstances of use and reception (Allan, 2015; Rahman, 2012).

2.2 Annotation Paradigms

Contextual swearing and minority language pose major challenges to simplistic judgments relying solely on phrasal units and general definitions (Pamungkas et al., 2023; Deas et al., 2023). Simple reminders of exceptions and rare cases are insufficient, as unrestricted context interpretation based on individual assumptions inevitably introduces biases (Rast, 2009). Educative annotation decisions regarding context must follow predefined instructions (Giunchiglia et al., 2017; Röttger et al., 2021). Descriptive data annotation embraces subjectivity to gain insights into diverse viewpoints but faces challenges in effectively eliciting, representing, and modeling those viewpoints (Röttger et al., 2021; Alexeeva et al., 2023). Prescriptive data annotation standardizes annotated features to provide consistent views of targeted language usages but risks overlooking some acceptable interpretations (Röttger et al., 2021; Ruggeri et al., 2023). Mitigating the potential deficiency of prescriptive annotation paradigms is a major issue in establishing this new benchmark.

2.3 Studies-Driven Definition for Toxic Language

Toxic language, a broader term than hate speech, refers to harm-inflicting expressions (Buell, 1998; Radfar et al., 2020; Baheti et al., 2021). Hate speech, characterized by emotional and direct aggression towards targets (Gelber, 2019; Elsherief et al., 2018), is a manifestation of toxic language rather than being equivalent to it (Fortuna et al., 2020). Treating toxicity and hatred separately avoids potential confusion arising from treating them as interchangeable concepts. Offensiveness and toxicity in language are characterized by their capacity to evoke negative reactions, distinct from mere swear word usage (Legroski, 2018), and are tied to linguistic politeness and social decorum (Archard, 2014), emphasizing the intention to denigrate rather than actual harm inflicted (Archard, 2008). Aggressiveness, while fundamental to dominating behavior (Kacelnik and Norris, 1998), differs from outward toxicity that adversely impacts others. Aggressive components may contribute to offensive speech only when coupled with explicit intents to cause harm or distress (Stokes and Cox, 1970). In short, toxic offensive language is the lan-

guage that shows explicit aggression towards others. Separating language aggression from language intent aims to direct human judgment in annotation onto relevant textual features, avoiding biases and improving agreement by not erroneously marking provocative but ultimately inoffensive speech as inappropriate.

3 Methodology

Two components need to be assessed to determine toxicity: the direction of language intent (DI) and the presence of aggression (AG). DI has two labels: 1 for explicitly targeting other people and 0 for other cases. AG has three labels: 0 for non-aggressive, 1 for mildly aggressive, and 2 for intensely aggressive. A piece of text is categorized as **toxic or offensive if and only if it is labeled as 1 for DI and either 1 or 2 for AG**. The logic form is shown as follows:

$$\forall x (\text{Toxic}(x) \iff (\text{DI}(x) = 1) \wedge (\text{AG}(x) = 1 \vee \text{AG}(x) = 2))$$

3.1 Annotation Criteria

Direction of Intent (DI) indicates whether the language is directed externally (label 1) or not (label 0). Text segments receive a label of 1 if they directly refer to or address a specific person or group using second-person pronouns, proper nouns, or clear contextual references that signal an interpersonal attack or criticism. Text segments receive a label of 0 if the statements implicate others more implicitly, as is common with ironic expressions, or focus primarily on the speaker themselves. This simplified dichotomization aims to delineate clear instances of directive aggressive speech from more ambiguous cases. Since a tweet may contain multiple sentences with shifting targets, keeping disagreement in annotations is necessary for overlooking possible interpretations.

Aggression (AG) is annotated by categorizing negative, rude, or hostile attitudes into three levels: non-aggression (label 0, score 0), mild aggression (label 1, score 1), and intense aggression (label 2, score interval $(1, \infty)$). Table 1 provides a relative score reference for categorizing and quantifying linguistic aggression across lexical, syntactic, and discourse levels. Linguistic items are classified as aggressive items (AI) that independently convey aggression or aggression catalyzers (AC) that intensify aggression but are not inherently aggressive.

Level	Item	Category	Example
Lexical	Aggressive Noun Phrase and Determiner Phrase	<i>Aggressive Item</i>	Stereotyped noun phrase/determiner phrase (nigga, chingchong, <i>etc.</i>), bitch, shit, dumbass, <i>etc.</i>
Lexical	Aggressive Verb Phrase	<i>Aggressive Item</i>	fuck, hate, <i>etc.</i>
Lexical	Aggressive Adjective Phrase	<i>Aggressive Item</i>	retarded, psycho, stupid, <i>etc.</i>
Lexical	Aggressive Adverb Phrase	Aggression Catalyzer	fucking, <i>etc.</i>
Syntactic	Strong Expression	Aggression Catalyzer	should, must, definitely, <i>etc.</i>
Syntactic	Rhetorical Question	Aggression Catalyzer	Doesn't everyone feel the same? <i>etc.</i>
Syntactic	Imperative	Aggression Catalyzer	Shut the door, <i>etc.</i>
Discourse	Ironic Expression	Aggression Catalyzer	Clear as mud, <i>etc.</i>
Discourse	False Construct	<i>Aggressive Item</i> or Aggression Catalyzer	Those are people who only believe in flat earth, <i>etc.</i>
Discourse	Controversial Content	<i>Aggressive Item</i>	Inappropriate Content (adult, religious, <i>etc.</i>), jeering at others' mistakes or misfortunes, <i>etc.</i>

Table 1: **Relative Aggression Scoring Reference:** Assigns numerical values for aggressive speech: 1 point for Aggressive Items (overtly toxic statements) and 0.5 points for Aggression Catalyzers (toxicity booster). The false construct will be an exception.

AIs (e.g., slurs, vulgarities, inflammatory content) are weighted 1 point, and ACs (e.g., emphatic language, rhetorical questions, imperatives, ironic expressions) 0.5 points. False constructs, which lead to flawed evaluations or unfair treatment, become AIs when paired with ACs but are still worth 0.5 points. In calculating the relative aggression score, each unique linguistic item should be counted only once, as including multiple items from one category does not typically increase aggressiveness. Lastly, to reduce the risk of overlooking possibilities, we encouraged annotators to keep different interpretations of ACs, as they are usually more implicit and open to various interpretations.

3.2 Case Study

The following two case studies will demonstrate how our proposed annotation guidelines help mitigate biases by providing a clear framework for assessing the direction of intent (DI) and the level of aggression (AG).

Example (1) demonstrates casual language usage: "And apparently I'm committed to going to a new level since I used the key. Well, FUCK. Curiosity killed the Cat(hy)" (Barbieri et al., 2020). We apply our annotation criteria to assess its toxicity. This example includes the aggressive verb phrase F**K, categorized as an aggressive item (AI), leading to an aggression score of 1, which indicates mild aggression. However, since the statement does not explicitly target any individual, its DI (Directed Insult) is labeled as 0. According to

our criteria, a text is considered toxic or offensive only if it has a DI label of 1 and an AG label of either 1 or 2. Thus, example (1) is classified as non-toxic.

Example (2) illustrates the use of non-mainstream African American English: "I ain't never seen a bitch so obsessed with they nigga😂. I'm obsessed with mine😑" (Davidson et al., 2017). This example contains two aggressive noun phrases ("b***h" and "n***a"), both categorized as AI. However, according to our guidelines, each unique linguistic item is counted only once when calculating the aggression score, resulting in an aggression score of 1, indicating mild aggression. Additionally, as the statement does not explicitly target another individual, its DI is labeled as 0. Despite the use of aggressive language, the absence of explicit targeting results in a non-toxic classification based on our annotation criteria.

3.3 Human Annotation

Two separate annotation processes were conducted, one with predefined criteria and one without. For the non-criteria-based human annotation, two annotators were given the question prompt, "Is the tweet toxic or offensive? If toxic or offensive, label 1; if it is not, label 0." allow unrestricted subjectivity, following the descriptive data annotation paradigm. To examine the reliability of the original annotation, two annotators with academic backgrounds were chosen to resemble the diverse

and unspecified backgrounds of CrowdFlower(CF) workers who were randomly employed and coded for Davidson et al., 2017. The first annotator was a graduate marketing student familiar with internet culture but with no formal linguistic knowledge. The second was a graduate linguistics student with sufficient linguistic knowledge and socio-linguistic practices. Choosing annotators this way allowed evaluation of the reliability between the original and the descriptive data annotation under similar annotation conditions. The annotation with criteria was conducted by two linguistics graduate students who were trained with prescriptive instructions as presented in Appendix A . Please find more information about annotators and more details about the annotation process in Appendix B.

3.4 LLM Annotation

Leveraging in-context learning is a promising approach to mitigate various learning biases while ensuring low-cost and highly generalizable processing (Lampinen et al., 2022; Margatina et al., 2023; Coda-Forno et al., 2023). Few-shot learning enables language models to rapidly adapt to new downstream tasks by analyzing a small set of relevant examples or interactions to discern expected outputs without extensive retraining (Gao et al., 2020; Perez et al., 2021; Mahabadi et al., 2022).

This study uses GPT-3.5-turbo and GPT-4 to generate prototypical responses with proposed criteria prompts. GPT-3.5’s extensive architecture allows it to grasp and generate contextually relevant responses with limited input (Yang et al., 2021). GPT-4 further enhances this capability due to its even more extensive training and sophisticated design (OpenAI, 2023). We accessed both models via APIs to use small amounts of task-specific instruction to adapt to this task. Unlabeled data were processed with carefully constructed prompts to generate annotations consistent with pre-established formats. For descriptive LLM annotation, the question prompt used for human annotation was directly entered. For criteria-based LLM annotation, prompts were designed separately for the direction of intent, aggression recognition, and aggression scoring. The direction of intent prompt used general prescriptive instructions, while the aggression level prompt combined prescriptive instructions with few-shot examples sourced from the ‘AI’ and ‘AC’ categories to demonstrate specific scenarios. Given the subjective nature of aggression, includ-

Pair	CK	AC1	Agr.%
<i>Descriptive</i>			
1T & 2T	0.5172	0.5094	76.50
<i>Prescriptive & Descriptive</i>			
1T & 1T_C	0.3000	0.2406	66.75
2T & 1T_C	0.3889	0.3718	75.75
1T & 2T_C	0.2883	0.2229	66.25
2T & 2T_C	0.3966	0.3769	76.25
<i>Prescriptive</i>			
1AG_C & 2AG_C	0.8422	0.8419	90.75
1DI_C & 2DI_C	0.5913	0.5908	91.50
1T_C & 2T_C	0.7487	0.7486	92.50

Table 2: **Inter-Annotator Reliability Evaluation for Prescriptive and Descriptive Annotations:** 1T denotes descriptive toxicity, marketing student; 2T denotes descriptive toxicity, linguistics student; 1AG_C denotes prescriptive aggression, Annotator 1; 2AG_C denotes prescriptive aggression, Annotator 2; 1DI_C denotes prescriptive intent direction, Annotator 1; 2DI_C denotes prescriptive intent direction, Annotator 2; 1T_C denotes prescriptive toxicity, Annotator 1; 2T_C denotes prescriptive toxicity, Annotator 2

ing some examples in the latter prompt was crucial for ensuring some uniformity in annotations. Additionally, the challenge of neurotoxic degeneration is tackled by employing a method similar to Instruction Augmentation (INST) (Prabhumoye et al., 2023). We divided the aggression level prompt into two sections: one for assessing language use and another for aggression scoring. This division adheres to INST principles, enhancing the clarity and precision of instructional prompts for saving effects in cleaning the outcomes.

4 Data Analysis

We randomly collected 400 tweets from the Offensive and Hate Speech dataset of the Davidson 2017 dataset (Davidson et al., 2017). This dataset contains a high frequency of various types of offensive language and non-mainstream English. We chose this dataset because its dense toxic content and casual language use make it relatively straightforward for both human annotators and language models to process. The prevalence of clear toxic content reduces potential confusion and ambiguity that could skew the analysis.

4.1 Inter-Annotator Agreement and Validation Analysis

Confusion matrices for all annotations are listed in Appendix C, and the distributions are displayed in Appendix D. For a comprehensive evaluation of annotator consistency, we calculated Cohen’s

Pair	CK	AC1	Agr. %
1T & Davidson et al., 2017	-0.0475	-0.2552	51.25
2T & Davidson et al., 2017	-0.0566	-0.1742	62.25
1T_C & Davidson et al., 2017	-0.0884	-0.1237	75.00
2T_C & Davidson et al., 2017	-0.0405	-0.0698	77.00

Table 3: Inter-annotator Reliability Evaluation on prescriptive, descriptive, and original annotation.

Kappa (CK) (McHugh, 2012) and Gwet’s AC1 (AC1)(Cicchetti, 1976), as detailed in Table 2. Initially, we assessed the inter-annotator reliability for both our annotations without criteria and those from Davidson et al., 2017, displayed in Table 3. Gwet’s AC1 can help avoid the paradoxical behavior and biased estimates associated with Cohen’s Kappa, especially in situations of high agreement and prevalence (Zec et al., 2017).

According to Table 2, incorporating specific criteria in the annotation process significantly enhances consistency and agreement between raters. This conclusion is supported by the larger positive values of trinary metrics for with-criteria pairs compared to without-criteria pairs and with-without-criteria pairs. Cohen’s Kappa and Gwet’s AC1 values, which adjust for chance agreement, indicate only moderate agreement without criteria. However, these values markedly increased when criteria were applied, as the first and last pairs approached near-perfect agreement levels. This underscores the critical role of well-defined criteria in enhancing reliability and validity of qualitative assessments. Interestingly, the reliability evaluations for with-without-criteria pairs are even lower than without-criteria pairs, suggesting the annotation logic for the two annotation types are entirely different.

Unlike our annotations, the comparison with the original annotations presents contrasting results in Table 3. Cohen’s Kappa and Gwet’s AC1 values are negative across all comparisons, suggesting a level of disagreement more pronounced than random chance. This also indicates underlying distinctions in how the annotations were carried out, and the fact that the majority vote labels they used for the final label were not from the same annotator could be a reason why reliability tests exhibit so much difference. These statistics starkly contrast the earlier findings where criteria application resulted in a near-perfect agreement for specific pairs. Although the agreement percentages showed some surface agreement, they do not align with the deeper discordance indicated by the negative Cohen’s Kappa

and Gwet’s AC1 values. As a result, prescriptive data annotations (1T_C, 2T_C) show higher reliability compared to descriptive data annotations (1T, 2T). Prescriptive data annotation paradigms are more appropriate for this task. This discrepancy highlights the complexities in achieving inter-rater reliability and the need to thoroughly review annotation guidelines and processes to understand and rectify the significant misalignments.

4.2 Validation and Agreement Analysis of Human and GPT Annotations

As Cohen’s Kappa and Gwet’s AC1 were created to assess inter-rater reliability between human annotators, directly applying them to evaluate agreement between machine and human annotations may not be entirely apt (Popović and Belz, 2021). While primarily intended for only human judgment scenarios, we include evaluations using these metrics when comparing GPT model predictions and human labels since dedicated methods for assessing machine-human agreement have yet to be established. We analyzed the concordance between human annotations and those generated by GPT models, namely GPT-4 (OpenAI, 2023) and GPT-3.5 (OpenAI, 2022), across two annotation categories. The trinary evaluations in Table 4 demonstrate reasonable consistency and agreement between human annotations and those from GPT-3.5 and GPT-4. Without prompted criteria, GPT-3.5 slightly outperforms GPT-4 in both agreement and reliability, but refining the prompts enabled more effective and reliable synergy between automated toxicity analysis and human-like interpretation. Using the proposed criteria significantly improved the alignment with human judgment for both models, especially for GPT-4 annotations. Inter-rater reliability Under criteria-based scenarios, GPT-4 annotations showed comparable agreement and consistent inter-rater reliability. The reliability statistics show that GPT annotations have even higher agreement and consistency than the original human annotations and without-criteria human annotations following

Pair	CK	AC1	Agr. %	Pair	CK	AC1	Agr. %
<i>Without Criteria</i>							
1T & G4T	0.2030	0.0685	62.75	1T & G3T	0.3149	0.2532	67.50
2T & G4T	0.2819	0.2190	73.75	2T & G3T	0.3534	0.3331	74.50
<i>With Criteria</i>							
1DI_C & G4DI_C	0.3376	0.3361	87.00	1DI_C & G3DI_C	0.1999	0.1799	87.75
2DI_C & G4DI_C	0.5647	0.5646	92.25	2DI_C & G3DI_C	0.2820	0.2704	90.25
1AG_C & G4AG_C	0.3460	0.3016	62.5	1AG_C & G3AG_C	0.2813	0.2605	59.25
2AG_C & G4AG_C	0.3849	0.3565	66.5	2AG_C & G3AG_C	0.2700	0.2588	60.0
1T_C & G4T_C	0.5299	0.5282	87.00	1T_C & G3T_C	0.4013	0.3887	85.5
2T_C & G4T_C	0.6103	0.6094	89.50	2T_C & G3T_C	0.4015	0.3910	86.0

Table 4: **Inter-Annotator Reliability Evaluation of GPT Annotations and Human Annotations:** G4T denotes descriptive toxicity, GPT-4; G3T denotes descriptive toxicity, GPT-3.5-turbo; G4DI_C denotes prescriptive intent direction, GPT-4; G4AG_C denotes prescriptive aggression, GPT-4; G4T_C denotes prescriptive toxicity, GPT-4; G3DI_C denotes prescriptive intent direction, GPT-3-turbo; G3AG_C denotes prescriptive aggression, GPT-3.5-turbo; G3T_C denotes prescriptive toxicity, GPT-3.5-turbo

Model (Fine-Tuning Data)	DI (F1)	AG (F1)	T (F1)
RoBERTa-base (Davidson et al., 2017)	-	-	0.912
DeBERTa-base (Davidson et al., 2017)	-	-	0.908
RoBERTa-base (G3P)	0.894	0.656	-
DeBERTa-base (G3P)	0.913	0.715	-
RoBERTa-base (G4P)	0.927	0.849	-
DeBERTa-base (G4P)	0.925	0.825	-

Table 5: Learning Performance for BERT models Fine-tuned on Davidson et al., 2017 baseline and GPT-annotated Datasets with Macro-averaged F1

the descriptive paradigm. The established criteria improved accuracy. Additionally, GPT-4 outperformed GPT-3.5 on this task. This suggests an aptitude for criteria-based analysis. After implementing the proposed criteria, these notable improvements demonstrate that prescriptive data annotation instructions can help researchers overcome the lack of human annotator resources.

5 Experiments

The experiment settings involve fine-tuning two models, RoBERTa-base with approximately 125 million parameters (Liu et al., 2019) and DeBERTa-base with approximately 139 million parameters (He et al., 2021), using a training batch size of 8 and an evaluation batch size of 16 with $5e-5$ learning rate. The models are trained for 3 epochs, with the dataset split into 90% for training and 10% for testing. To stabilize training, a learning rate warmup strategy is employed with 500 warmup steps. Weight decay regularization with a value of 0.01 is applied to prevent overfitting by encouraging smaller weights. Two datasets were used in this study. The baseline models were fine-tuned

on 2,438 tweets from the Davidson 2017 dataset (Davidson et al., 2017), excluding 400 pieces used in statistical analysis. In comparison, a 1,942-piece dataset was compiled for prescriptive LLM annotations, consisting of 295 Reddit posts in African American English (Deas et al., 2023), 341 tweets from OLID (Zampieri et al., 2019), 311 tweets from the offensive and hate speech dataset (Davidson et al., 2017), and 1,000 tweets from Hateval (Basile et al., 2019). The combination of different datasets helps mitigate extrusive language features, while the inclusion of diverse social media platforms (e.g., Reddit, Twitter) facilitates robust exposure to various language types and dialects. Previous studies and empirical observations suggest that larger datasets, particularly those with language types similar to the target application, tend to lead to higher performance in language models (Sahlgren and Lenci, 2016; Linjordet and Balog, 2019; Kaplan et al., 2020). Therefore, the Davidson 2017 dataset, with its size and domain relevance advantages, would likely enable superior performance compared to the smaller, more complex 1,942-piece dataset.

Model (Fine-Tuning Data)			1T	2T		
RoBERTa-base (Davidson et al., 2017)			0.379	0.665		
DeBERTa-base (Davidson et al., 2017)			0.379	0.531		
	1DI_C	2DI_C	1AG_C	2AG_C	1T_C	2T_C
RoBERTa-base (Davidson et al., 2017)	-	-	-	-	0.728	0.742
DeBERTa-base (Davidson et al., 2017)	-	-	-	-	0.728	0.742
RoBERTa-base (G3P)	0.828	0.867	0.597	0.572	0.806	0.819
DeBERTa-base (G3P)	0.839	0.877	0.525	0.558	0.793	0.811
RoBERTa-base (G4P)	0.850	0.889	0.389	0.446	0.837	0.859
DeBERTa-base (G4P)	0.879	0.908	0.383	0.441	0.817	0.839

Table 6: Macro-averaged F1 Scores of BERT models fine-tuned on Davidson et al., 2017 baseline and GPT-annotated data in Comparison with Human Annotations

5.1 Result Analysis and Discussion

As shown in Table 5, when fine-tuned on different datasets, DeBERTa-base slightly outperforms RoBERTa-base on the baseline dataset, achieving macro F1 scores of 0.908 and 0.912, respectively. However, RoBERTa-base achieves higher accuracy in prescriptive Aggression (AG) and prescriptive Direction of Intent (DI) when trained on GPT-annotated datasets (G3P² and G4P³). RoBERTa-base achieves macro F1 scores of 0.894 and 0.656 for DI and AG, respectively, on the G3P dataset and 0.927 and 0.849 on the G4P dataset. All experiments were conducted using an NVIDIA A100 GPU. Macro-F1 scores in Table 6 indicate that fine-tuned models align well with human annotations in identifying language intent (1DI_C and 2DI_C) but struggle more with aggression classifications (1AG_C and 2AG_C). When fine-tuned on the baseline dataset, BERT models moderately agree with human toxicity annotations (1T and 2T), with macro F1 scores of 0.379 for 1T and 0.665 and 0.531 for 2T using RoBERTa-base and DeBERTa-base, respectively. Notably, criteria-based auto-annotations improve model performance, with higher agreement rates using the G4P dataset. Models fine-tuned on G4P annotations achieved lower macro F1 scores for aggression (0.389 and 0.446 for 1AG_C and 2AG_C using RoBERTa-base) but higher macro F1 scores for toxicity (0.837 and 0.859 for 1T_C and 2T_C using RoBERTa-base).

These results suggest that GPT-4’s annotations may not have captured the features needed to distinguish between mild and intense aggression. Still, they did exhibit features that differentiate non-

aggressive from aggressive content. The similar and higher macro F1 scores for toxicity in models fine-tuned on G3P and G4P (ranging from 0.793 to 0.859) compared to baselines demonstrate the effectiveness of using properly-prompted LLMs over random human annotators. Despite improvements, fine-tuned BERT models still lag behind prescriptive human annotators and prescriptively-prompted LLM annotations, possibly due to small dataset sizes. This result contradicts the previous hypothesis that the baseline dataset with a much larger size and more uniform language patterns would help small models outperform LLM annotations; instead, it strongly suggests the robustness of models fine-tuned on prescriptively annotated data.

6 Conclusion

In conclusion, this study improves offensive language detection by introducing a prescriptive annotation benchmark that separately evaluates intent and aggression, reducing bias and preserving language diversity. Our analysis demonstrates that LLMs, guided by few-shot learning and clear criteria, can identify annotation errors in casual and non-mainstream language, offering better reliability than previous studies. The proposed framework also improves BERT’s performance on small, complex datasets, outperforming baselines in resource-limited scenarios. These findings highlight the efficiency of this approach in optimizing data use and adapting toxic content moderation systems to diverse language patterns, even with limited annotation resources.

Limitations

First of all, aggressive expression classifications are not definitive. There is room for different interpre-

²1,942-piece set annotated by GPT-3.5-turbo with proposed criteria

³1,942-piece set annotated by GPT-4 with proposed criteria

tations to mitigate the risk of over-generalization associated with prescriptive annotation. What constitutes a specific category of aggression could shift over time as cultural norms and language use evolve. Additionally, it can sometimes be difficult to precisely categorize certain expressions of aggression due to variations in language, influences from popular culture, and other contextual factors. The following criteria only try to grasp a more objective overview of aggression, which does not intend to rule out all subjectivity. Putting values on categories assesses the functional diversity of different language components, providing a more precise evaluation of the aggression level. However, in certain instances, merely adding more terms from a single category can decrease the perceived aggression. This is because excessive repetition of similar aggressive language might come across as impotent rage, reducing the overall impact of the aggression expressed.

We identified some limitations that are important for guiding future research. While prescriptive annotation paradigms may better identify uniform patterns, they risk overlooking meaningful interpretations not yet recognized by linguists and social scientists. The proposed criteria account for variations in English, but their practical application relies heavily on annotators' language knowledge. The dynamic nature of internet language poses additional challenges for human coders to accurately comprehend tweets, as no annotators can fully grasp the breadth of English online language, let alone code-switching usages by multilingual users. On the other hand, annotators lacking contextual understanding of in-group language may erroneously analyze utterances meant to promote within-community comprehensibility, a limitation challenging to resolve through improved annotation design. In contrast, LLMs demonstrate an advantage in aggregating insights from considerably larger data sources. Therefore, determining approaches for incorporating LLMs in detection alongside human rationale remains an important direction for further research.

Furthermore, the scope of human annotation within our dataset could be expanded. Human annotation of a dense toxicity corpus reveals high agreement; however, corpora containing more implicit cultural-related expressions would likely yield lower agreement rates. So, the human agreement in this research is only a reference, not a

solid upper bound. Although we relied on a significant amount of human input, the complexities and nuances of offensive language suggest that a broader and more diverse set of human annotations could enhance the model's understanding and accuracy. Another limitation lies in the size of our auto-annotated dataset. Additionally, there is room for improvement in the performance of smaller models on the automatically generated dataset. Open-source LLMs could be possible substitutes. Exploring different configurations, experimenting with various model architectures, and further tuning could enhance performance.

Acknowledgement

I would like to sincerely thank Hao Yu and Kedi Mo from Teachers College, Columbia University, and Jiaqi Wang from the Olin Business School of Washington University in St. Louis for their invaluable help with the annotations for this project. Their hard work and dedication were crucial to its completion. Special thanks are due to Nicholas Deas and Professor Kathleen McKeown from the Fu Foundation School of Engineering and Applied Science at Columbia University, who guided me in conducting research and designing the research framework. I would also like to express my gratitude to Dr. Howard A. Williams and Dr. Payment Vafae from Teachers College, Columbia University, whose insightful brainstorming sessions and valuable discussions helped shape the boundary between aggressive language and toxic language, providing further insight into addressing toxic language. Their guidance was instrumental in refining the framework for this study. I would also like to extend my sincere thanks to Dr. Erik Voss from Teachers College, Columbia University, my MA project advisor, for his unwavering support and guidance throughout the drafting of this research. Finally, I would like to thank the reviewers for their thoughtful suggestions, which helped improve the clarity of the paper.

References

- Maria Alexeeva, Caroline Hyland, Keith Alcock, Allegra Argent Beal Cohen, Hubert Kanyamahanga, Isaac Kobby Anni, and Mihai Surdeanu. 2023. [Annotating and training for population subjective views](#). In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.
- Keith Allan. 2015. When is a slur not a slur? the use of

698	nigger in ‘pulp fiction’. <i>Language Sciences</i> , 52:187–199.	Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of african american language bias in natural language generation .	750
699			751
700	David Archard. 2008. Disgust, offensiveness and the law. <i>Journal of Applied Philosophy</i> , 25(4):314–321.		752
701			753
702	David Archard. 2014. Insults, free speech and offensiveness. <i>Journal of Applied Philosophy</i> , 31(2):127–141.	Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In <i>Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society</i> , pages 67–73.	754
703			755
704	Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark O. Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts . <i>ArXiv</i> , abs/2108.11830.		756
705			757
706			758
707			
708	Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In <i>Proceedings of Findings of EMNLP</i> .	Mai Elshierief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding-Royer. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media . In <i>International Conference on Web and Social Media</i> .	759
709			760
710			761
711			762
712			763
713	Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In <i>Proceedings of the 13th international workshop on semantic evaluation</i> , pages 54–63.	Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing</i> . Association for Computational Linguistics.	764
714			765
715			766
716			767
717			768
718			769
719			770
720	Lawrence Buell. 1998. Toxic discourse . <i>Critical Inquiry</i> , 24:639 – 665.	Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In <i>Proceedings of the 12th language resources and evaluation conference</i> , pages 6786–6794.	771
721			772
722	Domenic V Cicchetti. 1976. Assessing inter-rater reliability for rating scales: resolving some basic issues. <i>The British Journal of Psychiatry</i> , 129(5):452–456.		773
723			774
724			775
725	Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew Botvinick, Jane X. Wang, and Eric Schulz. 2023. Meta-in-context learning in large language models .	Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In <i>Proceedings of the international AAAI conference on web and social media</i> , volume 12.	777
726			778
727			779
728			780
729	Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes . <i>Transactions of the Association for Computational Linguistics</i> , 11:300–319.	Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. <i>arXiv preprint arXiv:2012.15723</i> .	781
730			782
731			783
732			784
733			785
734	Aida Mostafazadeh Davani, M. C. D’iaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations . <i>Transactions of the Association for Computational Linguistics</i> , 10:92–110.	Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey .	787
735			788
736			789
737			
738			790
739	Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In <i>Proceedings of the international AAAI conference on web and social media</i> , volume 11, pages 512–515.	Katharine Gelber. 2019. Terrorist-extremist speech and hate speech: Understanding the similarities and differences . <i>Ethical Theory and Moral Practice</i> , pages 1–16.	791
740			792
741			793
742			
743			794
744	Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum . In <i>Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)</i> , pages 11–20, Brussels, Belgium. Association for Computational Linguistics.	Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets . <i>ArXiv</i> , abs/1908.07898.	795
745			796
746			797
747			
748			798
749			799
		Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. 2017. Personal context modelling and annotation . <i>2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)</i> , pages 117–122.	800
			801
			802

803	Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention . In <i>International Conference on Learning Representations</i> .	OpenAI. 2022. Gpt-3.5: Language models are few-shot learners. https://openai.com/blog/gpt-3-5-update/ . Accessed: [Insert current date here].	858 859 860 861
807	Alejandro Kacelnik and Sasha Norris. 1998. Primacy of organising effects of testosterone . <i>Behavioral and Brain Sciences</i> , 21:365 – 365.	OpenAI. 2023. Gpt-4 technical report .	862
810	Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models .	Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Investigating the role of swear words in abusive language detection tasks. <i>Language Resources and Evaluation</i> , 57(1):155–188.	863 864 865 866
814	Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators’ demographic characteristics . In <i>Workshop on Abusive Language Online</i> .	Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. <i>Advances in neural information processing systems</i> , 34:11054–11070.	867 868 869 870
818	Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? <i>arXiv preprint arXiv:2204.02329</i> .	Maja Popović and Anya Belz. 2021. A reproduction study of an annotation-based human evaluation of mt outputs. Association for Computational Linguistics (ACL).	871 872 873 874
824	Marina Chiara Legroski. 2018. Offensiveness scale: how offensive is this expression? <i>Estudos Linguísticos (São Paulo. 1978)</i> , 47(1):169–180.	Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Adding instructions during pretraining: Effective way of controlling toxicity in language models . In <i>Conference of the European Chapter of the Association for Computational Linguistics</i> .	875 876 877 878 879 880
827	Trond Linjordet and Krisztian Balog. 2019. Impact of training dataset size on neural answer selection models. In <i>Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41</i> , pages 828–835. Springer.	Bahar Radfar, K. Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context . In <i>International Conference on Web and Social Media</i> .	881 882 883 884
833	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach . <i>CoRR</i> , abs/1907.11692.	Jacquelyn Rahman. 2012. The n word: Its history and use in the african american community. <i>Journal of English Linguistics</i> , 40(2):137–171.	885 886 887
838	Zoey Liu, Crystal Richardson, Richard J. Hatcher, and Emily Prudhommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology . In <i>Annual Meeting of the Association for Computational Linguistics</i> .	Erich H. Rast. 2009. Context and interpretation .	888
843	Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Marzieh Saeidi, Lambert Mathias, Veselin Stoyanov, and Majid Yazdani. 2022. Perfect: Prompt-free and efficient few-shot learning with language models. <i>arXiv preprint arXiv:2204.01172</i> .	Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. <i>arXiv preprint arXiv:2112.07475</i> .	889 890 891 892
848	Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. <i>arXiv preprint arXiv:2305.14264</i> .	Federico Ruggeri, Francesco Antici, Andrea Galassi, Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2023. On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection . In <i>Text2Story@ECIR</i> .	893 894 895 896 897
852	Mary L McHugh. 2012. Interrater reliability: the kappa statistic. <i>Biochemia medica</i> , 22(3):276–282.	Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. <i>arXiv preprint arXiv:1609.08293</i> .	898 899 900 901
854	Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. <i>arXiv preprint arXiv:2006.08328</i> .	Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. <i>arXiv preprint arXiv:1911.03891</i> .	902 903 904 905
855		Allen W Stokes and Lois M Cox. 1970. Aggressive man and aggressive beast. <i>BioScience</i> , 20(20):1092–1095.	906 907 908

909	Lucky Susanto, Musa Izzanardi Wijanarko, Prase-	– Aggression catalyzers / AC (.5 point):	957
910	tia Anugrah Pratama, Traci Hong, Ika Idris, Al-	are unaggressive themselves and func-	958
911	ham Fikri Aji, and Derry Wijaya. 2024. Indo-	tion to boost the aggressive level.	959
912	toxic2024: A demographically-enriched dataset of	– Expressions from the same item category	960
913	hate speech and toxicity types for indonesian lan-	only count once; for example, if there are	961
914	guage.	two different aggressive noun phrases,	962
915	Yi Chern Tan and Elisa Celis. 2019. Assessing social	the score will be one rather than two.	963
916	and intersectional biases in contextualized word rep-	– Override Rule: The overall relative ag-	964
917	resentations. <i>ArXiv</i> , abs/1911.01485.	gression score will be 0 if there is no	965
918	Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle	aggressive item.	966
919	Carvalho, Fabrício Benevenuto, and Thiago Alexan-	– SPECIAL CASE: False constructs are	967
920	dre Salgueiro Pardo. 2021. Contextual-lexicon ap-	non-aggressive. But when people pair	968
921	proach for abusive language detection. <i>arXiv preprint</i>	false constructs with other aggressive cat-	969
922	arXiv:2104.12265.	alyzers, they become aggressive items	970
923	Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei	(but with .5 point) and should be seen	971
924	Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021.	as aggression bases. For example, how	972
925	An empirical study of gpt-3 for few-shot knowledge-	come your people really believe in flat	973
926	based vqa. <i>ArXiv</i> , abs/2109.05014.	earth?	974
927	Marcos Zampieri, Shervin Malmasi, Preslav Nakov,	• Example contains examples of each item.	975
928	Sara Rosenthal, Noura Farra, and Ritesh Kumar.		
929	2019. Predicting the type and target of of-	Direction of Language Intent (External or Non-	976
930	fensive posts in social media. <i>arXiv preprint</i>	external) evaluates Whether the language targets	977
931	arXiv:1902.09666.	other(s) explicitly. The direction is decided regard-	978
932	Slavica Zec, Nicola Soriani, Rosanna Comoretto, and	ing the direction of aggression, which means even	979
933	Ileana Baldi. 2017. Suppl-1, m5: high agreement	statements about speakers’ selves could contain	980
934	and high prevalence: the paradox of cohen’s kappa.	aggression against others.	981
935	The open nursing journal , 11:211.		
936	A Annotator Codebook	B Annotator Surveys	982
937	A.1 General Definitions	Specialties	983
938	A list of short-cut definitions is presented in Table 7.	• Annotator 1 without criteria: Internet Market-	984
939	Please see the methodology for further validations.	ing & Data Analytics	985
940	A.2 Annotation Instruction for two Indicators	• Annotator 2 without criteria: Corpus Linguis-	986
941	Aggression will be assessed regarding every dis-	tics & Syntax	987
942	tinct negative, rude, or hostile attitude. Please see	• Annotator 1 with criteria: Semantics Analysis	988
943	Table 1 and general description below for more	& Syntax & Corpus Linguistics	989
944	information about specific language use. Compu-	• Annotator 2 with criteria: Socio-linguistics &	990
945	tation logic: If the score is less or equal to 1, the	Language Acquisition	991
946	aggression level will be 1. If the score exceeds		
947	1, the aggression level will be 2. Otherwise, the	Aside from mainstream English, are you	992
948	aggression level will be 0.	familiar with any regional dialects, sociolects,	993
949	• Level refers to the general linguistic category	or linguistic styles more common in minority	994
950	of each item.	communities and groups?	995
951	• Item name includes the names of aggression-	• Annotator 1 without criteria: Yes	996
952	related items.	• Annotator 2 without criteria: Yes	997
953	• Category refers to the category that indicates	• Annotator 1 with criteria: Yes	998
954	how the item is related to aggression.	• Annotator 2 with criteria: Yes	999
955	– Aggressive items / AI (1 point): are ag-		
956	gressive by themselves.		

Term	Definition
Aggression/Aggressiveness	Aggression in this context indicates hostile or rude attitudes, whether it involves readiness or not.
Aggressive	Being aggressive means showing hostile or rude attitudes, whether it involves readiness or not.
Offensiveness	General rudeness in a way that causes somebody to feel upset or annoyed because it shows a lack of respect.
Offensive	Being rude in a way that causes somebody to feel upset or annoyed because it shows a lack of respect.
External	Towards other people or parties.
Internal	Towards the self.
Construct	The mind-dependent object, namely ideas, perspectives, etc.
Inappropriate Language	Language uses that could have negative and unwanted impacts on people.
Biased Language	Biased Language contains obviously wrong or counterfactual expressions that target an individual or a group not limited to humans.
Offensive Language	Offensive Language shows intended aggressiveness toward others.
Hate Speech	Hate Speech is an offensive language of intense external aggressive intention with explicit targets rooted in explicit or implicit false constructs.

Table 7: Definitions of Terms

1000	Approximately how many hours did it take you to complete all the annotations assigned to you?	political stuff. I don't know too much about	1026
1001		politics.	1027
1002	• Annotator 1 without criteria: 4		
1003	• Annotator 2 without criteria: 4.5	• Annotator 2 with criteria: 4. Yes, I think AAE	1028
1004	• Annotator 1 with criteria: 5 (criteria-based	is not really an issue. The AG scoring guide	1029
1005	training) + 7 (annotation)	helps break things down to the word level. Ba-	1030
1006	• Annotator 2 with criteria: 5 (criteria-based	sically, it doesn't really matter if the phrases	1031
1007	training) + 8 (annotation)	are used differently or not; as long as they are	1032
1008	How confident are you in the accuracy of the	seen as aggressive by some people, they'll be	1033
1009	annotations you completed? (1-5)	taken as aggressive. But it really takes a lot of	1034
1010	• Annotator 1 without criteria: 2. No so confi-	time and effort just to highlight each aggres-	1035
1011	dent, many African American English I found	sive item and categorize the aggression. DI	1036
1012	hard to understand accurately	seemed pretty straightforward to me at first,	1037
1013	• Annotator 2 without criteria: 3. I am confi-	but after our group discussion, I realized there	1038
1014	dent about my annotations identifying explicit	could also be other interpretations.	1039
1015	toxic expressions and hate speech, but less		
1016	confident in others.	Looking back at your annotations after a	1040
1017	• Annotator 1 with criteria: 4.5. I'm pretty	month has passed, how did you feel about the	1041
1018	confident, though I'm not an African Amer-	quality and accuracy of the work you originally	1042
1019	ican English native speaker. I studied AAE	completed?	1043
1020	corpus before, so I consider myself familiar	• Annotator 1 without criteria: Still confused	1044
1021	with AAE. About that DI, sometimes I think	about many tweets.	1045
1022	it could go either way cause their tweets ain't	• Annotator 2 without criteria: There could be	1046
1023	just one sentence. For AG, the score generally	different interpretations. It's really about the	1047
1024	matches what I think about aggression. All	larger context.	1048
1025	in all, this dataset is easier than the one with	• Annotator 1 with criteria: Not really much in	1049
		terms of toxicity. DI's still kinda confusing in	1050
		a couple of cases.	1051

1052
1053
1054

- Annotator 2 with criteria: Basically the same as when I finished it up

C Confusion Matrices (Figure 2-5)

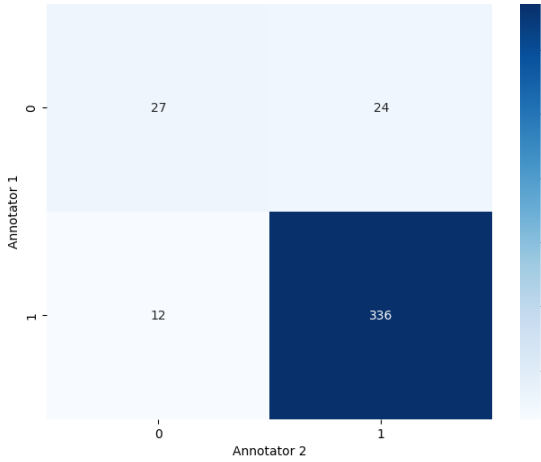


Figure 2: Confusion Matrix on Direction Intent Annotation

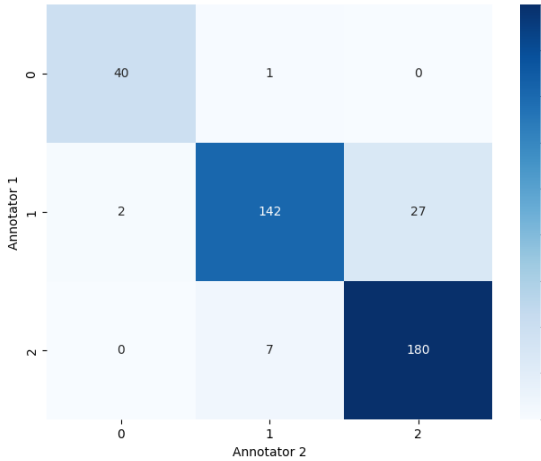


Figure 3: Confusion Matrix on Aggression Annotation

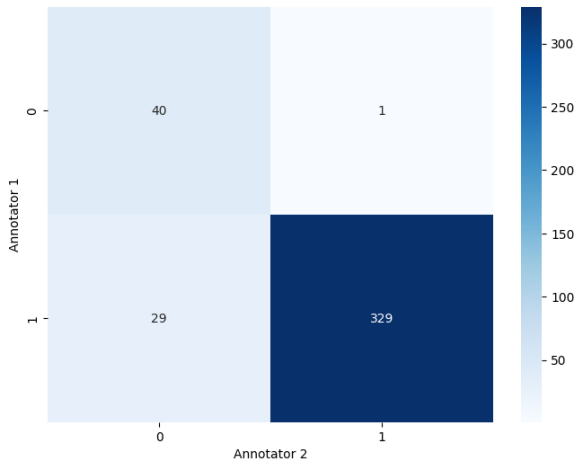


Figure 4: Confusion Matrix on Toxicity Annotation with Criteria

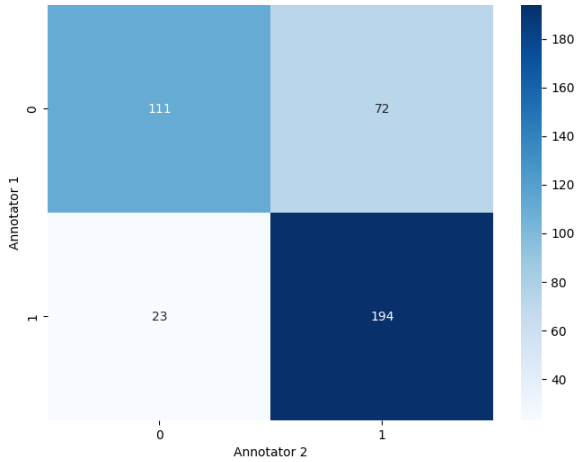


Figure 5: Confusion Matrix on Toxicity Annotation without Criteria

1055

D Annotation Distribution (Figure 6-9)

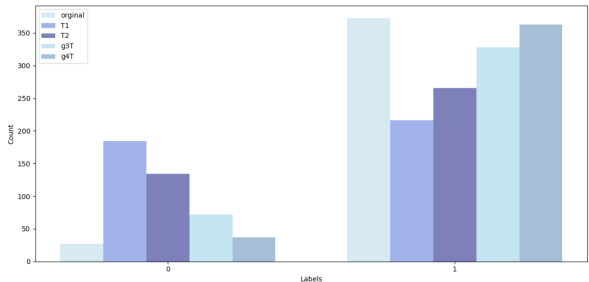


Figure 6: Distribution of Toxicity Annotation without Criteria

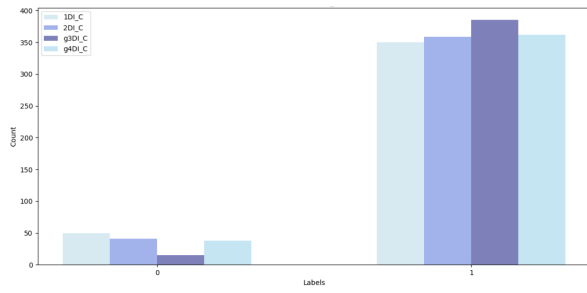


Figure 7: Distribution of Direction of Language Intent Annotation with Criteria

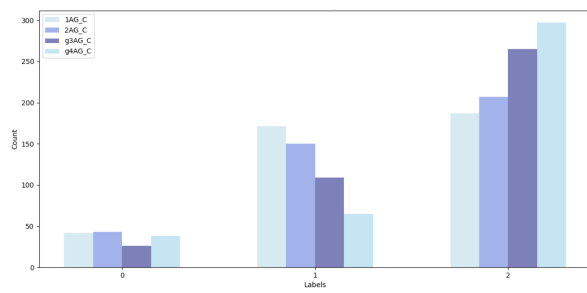


Figure 8: Distribution of Aggressive Level Annotation with Criteria

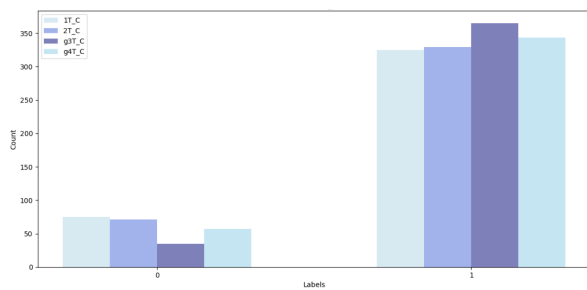


Figure 9: Distribution of Toxicity Annotation with Criteria

Classification of Buddhist Verses: The Efficacy and Limitations of Transformer-Based Models

Nikita Neveditsin¹, Ambuja Salgaonkar², Pawan Lingras¹, Vijay Mago³

¹Saint Mary's University, Halifax, Canada

²University of Mumbai, Mumbai, India

³York University, Toronto, Canada

Abstract

This study assesses the ability of machine learning to classify verses from Buddhist texts into two categories: Therigatha and Theragatha, attributed to female and male authors, respectively. It highlights the difficulties in data pre-processing and the use of Transformer-based models on Devanagari script due to limited vocabulary, demonstrating that simple statistical models can be equally effective. The research suggests areas for future exploration, provides the dataset for further study, and acknowledges existing limitations and challenges.

1 Introduction

The term "gāthā" (gatha) denotes a poetic meter primarily employed in legends and folklore, yet it is notably absent from the Vedas (Mukherjee, 1998). Gathas are popular in Maharashtra, India, where locals are familiar with the gathas of Tukaram (Tukaram, 2014). However, the earliest known reference to gathas appears in the Avesta, a Zoroastrian scripture compiled during the Sasanian Empire (224-651 BCE) (Hintze, 2002). The languages in which these ancient gathas were written have since become extinct. Consequently, interpreting them is challenging and necessitates reliance on extant languages that exhibit similar, yet distinctly different, structures.

This study examines two collections from the Buddhist canonical literature: *Theragathapali* and *Therigathapali*, which are, respectively, the line-wise utterances attributed to male and female saints. This literature is written in Pali, a language believed to be a mixture of Prakrit languages, closely related to the vernacular of the common people during the time of Siddhartha Gautama Buddha (circa 600 BCE).

The authorship of some gathas is debatable. Kumar (2016) observes that in Pali literature, authorship details are occasionally provided at the

beginning or end of the texts. However, not all authors considered it essential to include such information. In examining the authorship of the Therigatha, Findly (1999) suggests that the authorship of some verses may be doubtful, indicating that while some verses are traditionally attributed to the female saints themselves, others may have been composed or recited by different individuals, including the possibility of later attribution by compilers. This uncertainty in authorship challenges the straightforward attribution of these texts to the female saints they are associated with.

Nevertheless, studies demonstrate that the Therigathas differ from the Thera gathas. Blackstone (2013) argues that the Therigathas focus more on themes of overcoming suffering, societal constraints, and personal liberation. A study by Marques et al. (2021) confirms the uniqueness of topics in Therigatha.

Typically, a gatha is a two-line verse, although variations include verses comprising three or four lines. Figure 1 provides a sample two-line gatha in Devanagari script.

यो पुब्बे करणीयानि, पच्छा सो कातुमिच्छति।
सुखा सो धंसते ठाना, पच्छा च मनुतप्पति॥

Figure 1: Sample Gatha in Devanagari Script.

Banerjee (2017) suggests that translations of gathas influences the perception of these ancient texts. For example, the gatha from Figure 1 is translated by Bhikkhu (1998) as "Whoever wants to do later what he should have done first, falls away from the easeful state and later burns with remorse", while one of the contributors of this study translates the second line as "He destroys pleasure producing points and regrets later".

The abundance of Transformer-based models (Vaswani et al., 2017) and their proficiency across various domains (Fisher et al., 2023; Phatak et al.,

2024; Neveditsin et al., 2024), particularly in classification tasks (Munikaer et al., 2019; Kheiri and Karimi, 2023; Hartmann et al., 2023; Zielinski et al., 2023; Zaczynska et al., 2024), inspired us to conduct a study on their performance in classifying verses from low-resource Pali texts. While we acknowledge the debates around the authorship of some Therigatha verses, we deliberately avoid this discussion in our study due to the lack of definitive evidence regarding authorship. Consequently, we treat Therigatha verses as authored by female authors and Theragatha verses as authored by male authors.

The goal of this study is to determine whether Transformer-based models can outperform traditional machine learning models in the binary classification of the verses. We hypothesize that Transformer-based models, even when pretrained on languages other than Pali, can still identify patterns specific to each class. Additionally, we aim to assess the performance difference of these models when using Devanagari script versus Roman script. Through this investigation, we aim to highlight the challenges associated with this task and suggest directions for future research.

2 Related Work

Research on poetry classification in the Pali language using machine learning is scarce, however, insights can be drawn from related areas, including poetry classification in other languages, text classification in low-resource settings, and computational analysis of Pali texts.

One of the earliest studies in poetry classification is by Kao and Jurafsky (2012), who use logistic regression to examine stylistic and content features that distinguish professional poets from amateurs. The authors extract features related to diction, sound devices, affect, and imagery to identify elements contributing to poetic sophistication. Similarly, Pal and Patel (2020) classify Hindi poems using machine learning, providing insights into poetry classification in an Indo-Aryan language closely related to Pali. The authors employ classical models, such as Naïve Bayes, Random Forest, and SVM, achieving a maximum accuracy of 64% with Naïve Bayes, highlighting the challenges of poetry classification due to the morphological richness and varied sentence structures.

In the context of text classification for low-resource languages, recent research suggests that

cross-lingual models, such as XLM (Lample and Conneau, 2019), may sometimes offer performance gains compared to classic machine learning models like SVM or Naïve Bayes. For instance, Li et al. (2020) introduce a model called AgglutiFiT, fine-tuned from a cross-lingual pre-trained model (XLM-R), which significantly outperforms strong baselines in terms of accuracy.

Additionally, Alekseev et al. (2024) benchmark multilabel topic classification in the Kyrgyz language, evaluating several baseline models, including classical approaches and neural models like XLM-RoBERTa. Their findings indicate that the multilingual model XLM-RoBERTa outperforms classical models in terms of F1 score. However, transformer-based models do not always surpass traditional machine learning models for low-resource languages. For example, Lalhangmawii and Singh (2023) found that the SVM model achieved the highest accuracy (75%) on a sentiment classification task for the Mizo language, performing similarly to the XLM-RoBERTa model using a transfer learning approach.

Another method for handling low-resource languages is leveraging machine translation. Recent work by Kumar et al. (2024) provides valuable insights into sentiment classification for low-resource Indian languages using machine-translated datasets. The results highlight the potential of datasets translated with tools like Google Translate and indicate that models such as LSTM can effectively preserve sentiment by accounting for sequential patterns.

Focusing specifically on Pali texts, Zigmond (2021) conduct a computational analysis of the Pali Canon. The author uses computational text mining to examine various volumes of the Canon, extracting linguistic and thematic insights. By employing techniques such as k-means clustering and Principal Component Analysis, they reveal differences between older texts (Vinaya and Suttas) and later ones (Abhidhamma). The research also underscores the complexity of Pali language processing, including multiple word declensions, elisions, and compound formations.

3 Dataset

The dataset utilized in this study comprises the Thera and Theri gatha texts from the Khuddakanikaya volume of the Sutta-pitaka, which is the third part of the Buddhist canonical literature, Tip-

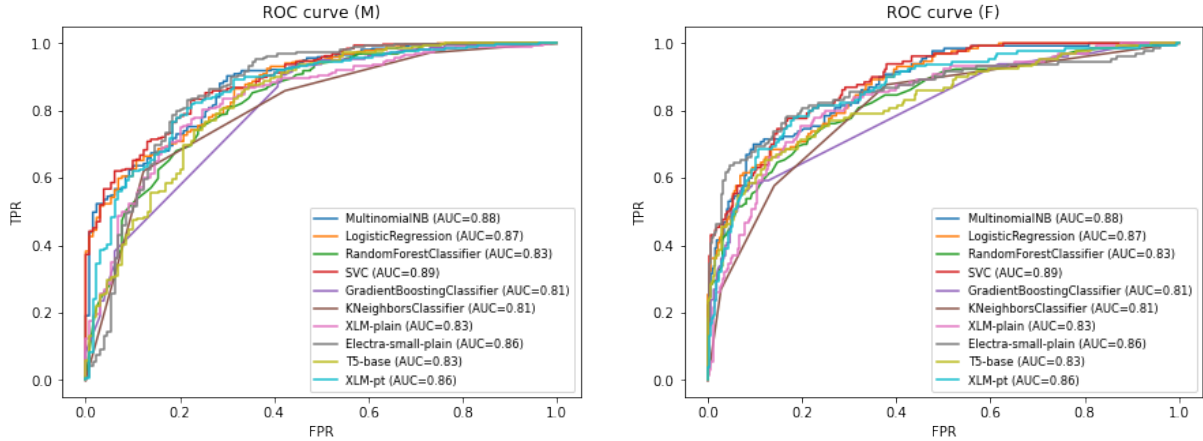


Figure 2: ROC curves for Roman script classification. Left: results for ‘M’ class (Theragatha); right: ‘F’ class (Therigatha). Multiple models are compared, with AUC scores indicating performance.

itaka¹. Each gatha is categorized into chapters based on the number of verses attributed to each author: single verses are compiled in the chapter named *Ekaka-nipaat*, meaning ‘collection of ones’, while chapters such as *Dukanipaat*—‘collection of twos’—contain texts with two verses from a single author, and so forth. The Theragatha consists of 1,288 verses spread across 21 chapters, whereas the Therigatha contains 524 verses distributed over 16 chapters, with all verses sequentially labeled within their respective compendiums.

To study the potential impact of script on the training of the classifier, both the Devanagari and Roman versions were used. The manual preprocessing involved several steps:

1. **Punctuation Handling:** We agreed on approaches to interpret punctuation marks, considering variations in their usage across different scripts.
2. **Text Completion:** This addresses instances of "*peyaala*" (or "*pe*"), which indicate a repetition of words or lines from previous parts of the text. Due to the lack of suitable computational linguistic tools for this task, matching the context of *peyaala* to find the appropriate text from earlier sections was conducted manually.
3. **Word Separation:** Ancient Indian languages feature notable word compounding and clubbing. Unlike Sanskrit, where the rules for word combination are relatively rigid, Pali allows more flexibility. This necessitates greater care in separating compounded words into their individual components. Due to the chal-

lenges in separating these combined words, we decided to work with the combined forms as they appear in the text.

After the manual preprocessing of the text, we encountered discrepancies in the counts of distinct words when tokenizing the verses by spaces. Assuming a one-to-one correspondence between tokens in the Devanagari and Roman scripts, a dictionary-based test was applied to identify these discrepancies. The test revealed several transliteration nuances. For instance, some symbols such as नौ and खौ in Devanagari are represented by two UTF-8 code points, which leads to confusion with symbols नं and खं, respectively. Another challenge was caused by complex compounding rules; for example, space-based tokenization ambiguously mapped the symbol मुनि to either ‘*muni*’ or ‘*munin*’, depending on the neighboring tokens (a one-to-many case). Similarly, both symbols न्ति and ति map to ‘*ti*’ in the Romanized script (a many-to-one case). These cases demonstrate that space-based tokenization may not adequately capture the nuances of these complex verses. For this study, we decided to exclude three nuanced verses from the Therigathas and sixteen nuanced verses from the Theragathas where we were unable to easily resolve the inconsistencies. This resulted in 1793 verses in our dataset². Table 1 presents the statistics on word distribution among the scripts.

4 Experiments and Results

The overall task can be defined as a binary classification problem with two categories: ‘M’ for Theragathas and ‘F’ for Therigathas. The dataset, divided

¹Digital version available here: <https://tipitaka.org/>

²<https://github.com/neveditsin/pali>

Statistic	Dev.	Rom.
Total Distinct Words	8787	8789
Female Unique Words	3145	3143
Male Unique Words	6548	6547
Only Female Words	2239	2242
Only Male Words	5642	5646
Common Words	906	901

Table 1: Word Distribution in Devanagari (Dev.) and Roman (Rom.) Texts.

by script type into Devanagari and Roman subsets, was split into training (75%) and test (25%) sets. Considering the dataset’s imbalance, we report key metrics such as ROC-AUC, Matthews Correlation Coefficient (MCC), as well as precision, recall, F1-scores, and average precision (AP) for both classes. We deliberately avoided sampling to address the imbalance due to the dataset’s small size. However, by providing a comprehensive set of metrics, we aim to give a detailed comparison of the models’ performance across different aspects.

First, we applied traditional machine learning models: Multinomial Naïve Bayes, Logistic Regression, Random Forest Classifier (RFC), Support Vector Classifier (SVC), Gradient Boost Classifier (GBC), and K-Nearest Neighbors Classifier (KNN), on the Roman script to classify gathas. Space tokenization and a TF-IDF matrix were used for all models except for the Multinomial Naïve Bayes, which served as a baseline model using simple count vectorization. The Multinomial Naïve Bayes assumes conditional independence of tokens and positional independence of features. Naïve Bayes can be optimal under certain circumstances, such as when the conditional independence assumption holds (Zhang, 2004). To assess whether transformer-based models could improve specific aspects of classification, such as precision and recall, we experimented with fine-tuning the following models: XLM (Lample and Conneau, 2019), XLM pre-trained additionally on our training corpus, T5-base (Raffel et al., 2023), and Electra-small (Clark et al., 2020). Figure 2 presents the classification results for the Roman script.

Similar experiments with the Devanagari script revealed that while transformer-based models underperformed relative to their counterparts in Roman script, the performance disparities among traditional models were minimal, as depicted in Figure 3. Additionally, our trials with a byte-level T5 (Xue et al., 2022) model yielded substantially lower performance (AUC 0.58 for Devanagari), which we attribute to its inability to effectively handle script-specific complexities, leading to its exclusion from

our study.

When investigating why transformer-based models exhibit inferior performance compared to classic machine learning algorithms, we analyzed the number of tokens generated by tokenizers for both Devanagari and Roman scripts in the test subsets. Table 2 presents the counts of unique tokens from the tokenizers applied to the test set. Our analysis revealed a strong correlation between the number of tokens and classification outcomes. This suggests that the underperformance of transformer-based models on the Devanagari script is attributed to significant information loss during tokenization with certain tokenizers.

	ByT5	OpenHathi	T5	XLM	Electra
Devanagari Tokens	54	1200	6	1208	60
Roman Tokens	44	-	748	1909	1313

Table 2: Unique Tokens in Test Subsets by Model

To address this issue, we opted to fine-tune OpenHathi-7B (Sarvam, 2024), a model based on Llama-2 (Hugo Touvron, 2023), specifically developed for Indo-Aryan languages. We utilized Low-Rank Adaptation (LoRA) (Hu et al., 2021) to adjust the model’s parameters, using the last token for classification purposes. Notably, even after fine-tuning, the OpenHathi model did not outperform the simpler XLM model.

Table 3 provides detailed classification results for the best performing models compared to Multinomial Naïve Bayes. Notably, OpenHathi exhibited the highest recall for the minority class among the evaluated models. However, a paired bootstrap test (Berg-Kirkpatrick et al., 2012) with 10^5 iterations indicated that this increase in recall is not statistically significant ($p = 0.08$).

Script	Class	Precision	Recall	F1	AP	AUC	MCC
Multinomial Naïve Bayes							
Devanagari	M	0.85	0.92	0.88	0.94	0.88	0.56
	F	0.75	0.60	0.67	0.78		
Roman	M	0.85	0.92	0.88	0.94	0.88	0.56
	F	0.75	0.60	0.67	0.78		
SVC							
Devanagari	M	0.86	0.88	0.87	0.95	0.89	0.53
	F	0.68	0.64	0.66	0.80		
Roman	M	0.86	0.88	0.87	0.95	0.89	0.53
	F	0.68	0.64	0.66	0.80		
XLM							
Devanagari	M	0.80	0.91	0.85	0.89	0.78	0.42
	F	0.68	0.45	0.54	0.64		
Roman	M	0.79	0.93	0.86	0.92	0.83	0.40
	F	0.71	0.39	0.50	0.67		
XLM with Pre-Training							
Devanagari	M	0.77	0.99	0.87	0.93	0.86	0.45
	F	0.95	0.28	0.44	0.75		
Roman	M	0.78	0.97	0.87	0.93	0.86	0.44
	F	0.83	0.35	0.49	0.75		
OpenHathi-7B							
Devanagari	M	0.85	0.68	0.76	0.86	0.76	0.36
	F	0.47	0.70	0.57	0.63		

Table 3: Detailed Results for Selected Models. Appendix A lists the hyperparameters used for

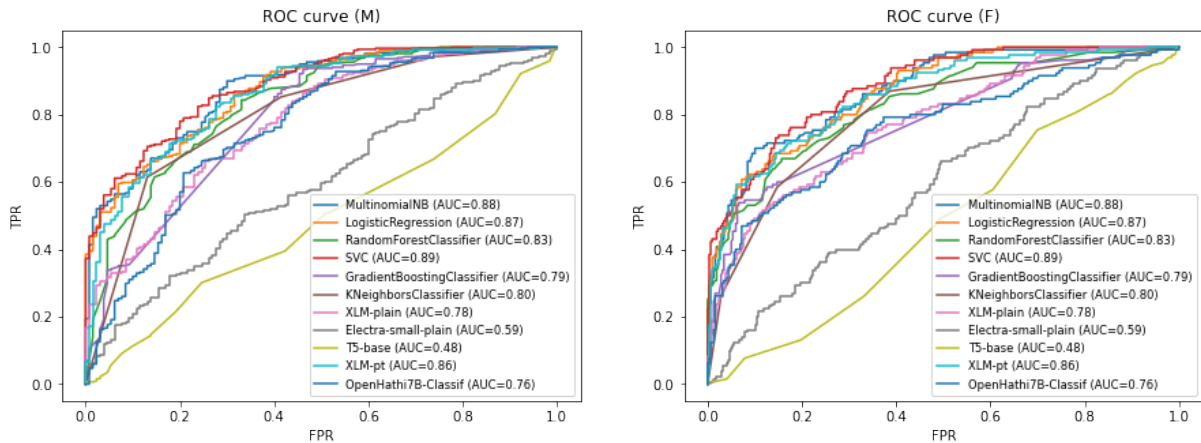


Figure 3: ROC curves for Devanagari script classification. Left: results for ‘M’ class (Theragatha); right: ‘F’ class (Therigatha). Multiple models are compared, with AUC scores indicating performance.

model training. Hyperparameters not listed are set to their default values in the scikit-learn library for classic machine learning models and in the Hugging Face Transformers library for transformer-based models.

Our attempt to employ the SHAP framework (Lundberg and Lee, 2017) on the best-performing models to explain their discrimination decisions did not reveal any specific features that contribute significantly to either of the classes.

5 Discussion and Further Research

The study highlights persistent challenges in using original, non-Romanized scripts with modern transformer-based models for classification tasks, primarily due to inadequate token coverage in the models’ vocabularies. Previous studies, such as the one by Maronikoulakis et al. (2021), showed that the compatibility of tokenizations is crucial in multilingual language models, discussing the importance of vocabulary size. More recently, Ali et al. (2024) confirmed that the choice of tokenizer significantly impacts a model’s downstream performance. They suggest that tokenizers not tailored to handle a variety of scripts can lead to inefficient tokenization, directly affecting model performance, and that larger vocabulary sizes are required for multilingual tokenizers compared to those designed for English only.

Although Romanized versions of the scripts enabled the use of a broader range of models, these models still did not surpass the performance of traditional machine learning algorithms. This outcome suggests that the employed models failed to identify any class-specific patterns within the

dataset, likely because these models lacked sufficiently relevant data during their pretraining stages. Notably, additional pre-training of the XLM model improved the AUC on the classification task, and a paired bootstrap test with 10^5 iterations confirmed the statistical significance of this improvement ($p < 0.05$).

Extended research is necessary for the authorship attribution task. Our next step is to identify Therigathas that are consistently misclassified by the majority of models and perform a detailed analysis of these cases. This includes annotating and analyzing specific gathas whose authorship is disputed by scholars. Statistical sampling to identify whether the differences between the Theri and Thera gathas are significant may help reveal if there are substantial distinctions between the two classes of gathas from a machine learning perspective. Additionally, compiling an extensive Pali corpus to pre-train a transformer model would enable us to experiment with its discriminatory abilities and its capability to generate novel gathas.

6 Limitations

First, our dataset is small and imbalanced, with only slightly over 10% of words shared between the Thera and Theri gathas. This low overlap might explain why classical machine learning algorithms were able to effectively discriminate between the classes, primarily by relying on words unique to specific classes.

The second limitation pertains to the existing transformer models, which often lack the comprehensive vocabulary necessary for thorough evaluation.

References

- Anton Alekseev, Sergey Nikolenko, and Gulnara Kabaeva. 2024. [Benchmarking multilabel topic classification in the kyrgyz language](#). In *Analysis of Images, Social Networks and Texts: 11th International Conference, AIST 2023, Yerevan, Armenia, September 28–30, 2023, Revised Selected Papers*, page 21–35, Berlin, Heidelberg. Springer-Verlag.
- Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. [Tokenizer choice for llm training: Negligible or crucial?](#) *Preprint*, arXiv:2310.08754.
- Supriya Banerjee. 2017. Ambapali's verse in therigatha: Trajectories and transformations. *Translation Today*, 11:15.
- Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. [An empirical investigation of statistical significance in NLP](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.
- Thanissaro Bhikkhu. 1998. [Harita \(2\) \(thag 3.15\)](#). Access to Insight (BCBS Edition), 4 August 2010.
- Kathryn R Blackstone. 2013. *Women in the Footsteps of the Buddha: Struggle for Liberation in the Therigatha*. Routledge.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [Electra: Pre-training text encoders as discriminators rather than generators](#). *Preprint*, arXiv:2003.10555.
- Ellison Banks Findly. 1999. [Women and the "arahant" issue in early pali literature](#). *Journal of Feminist Studies in Religion*, 15(1):57–76.
- Andrew Fisher, Matthew Maclaren Young, Doris Payer, Karen Pacheco, Chad Dubeau, and Vijay Mago. 2023. Automating detection of drug-related harms on social media: machine learning framework. *Journal of medical internet research*, 25:e43630.
- Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. [More than a feeling: Accuracy and application of sentiment analysis](#). *International Journal of Research in Marketing*, 40(1):75–87.
- Almut Hintze. 2002. On the literary structure of the older avesta. *Bulletin of the School of Oriental and African Studies*, 65(1):31–51.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Louis Martin et al. Hugo Touvron. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 8–17.
- Kiana Kheiri and Hamid Karimi. 2023. [Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning](#). *arXiv preprint arXiv:2307.10234*.
- Saurabh Kumar, Ranbir Sanasam, and Sukumar Nandi. 2024. [IndiSentiment140: Sentiment analysis dataset for Indian languages with emphasis on low-resource languages using machine translation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7689–7698, Mexico City, Mexico. Association for Computational Linguistics.
- S Vijitha Kumara. 2016. A chronological approach to the pali commentaries: with reference to the madhurattavilāsinī. *Sri Lanka International Journal of Buddhist Studies (SIJBS)*, 4:32–54.
- Mercy Lalthangmawii and Thoudam Doren Singh. 2023. [Sentiment analysis for the mizo language: A comparative study of classical machine learning and transfer learning approaches](#). In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 308–317, Goa University, Goa, India. NLP Association of India (NLP AI).
- Guillaume Lample and Alexis Conneau. 2019. [Cross-lingual language model pretraining](#). *Preprint*, arXiv:1901.07291.
- Xiuhong Li, Zhe Li, Jiabao Sheng, and Wushour Slam. 2020. [Low-resource text classification via cross-lingual language model fine-tuning](#). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 994–1005, Haikou, China. Chinese Information Processing Society of China.
- Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.
- Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. [Wine is not v i n. on the compatibility of tokenizations across languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2382–2399, Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Tattiane Yu Borges Marques et al. 2021. Therīgāthā, a primeira literatura feminina no budismo: possibilidades de diálogo com o erotismo na teopoética. *Mandrāgora*, 27(1):31–52.
- Sujit Mukherjee. 1998. *A dictionary of Indian literature: beginnings-1850*, volume 1. Orient Blackswan.
- Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. [Fine-grained sentiment classification using bert](#). In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5.
- Nikita Neveditsin, Pawan Lingras, and Vijay Mago. 2024. Clinical insights: A comprehensive review of language models in medicine. *arXiv preprint arXiv:2408.11735*.
- Kaushika Pal and Biraj. V. Patel. 2020. [Automatic multiclass document classification of hindi poems using machine learning techniques](#). In *2020 International Conference for Emerging Technology (INCET)*, pages 1–5.
- Atharva Phatak, Vijay K Mago, Ameeta Agrawal, Aravind Inbasekaran, and Philippe J Giabbanelli. 2024. Narrating causal graphs with large language models. *arXiv preprint arXiv:2403.07118*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Preprint*, arXiv:1910.10683.
- Sarvam. 2024. Announcingopenhathi series. <https://www.sarvam.ai/blog/announcing-openhathi-series>. Accessed: 2024-07-26.
- Sant Tukaram. 2014. *Tukaram Gatha: Enhanced by Rigved*, volume 1. Rigved Shenai.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. [Byt5: Towards a token-free future with pre-trained byte-to-byte models](#). *Preprint*, arXiv:2105.13626.
- Karolina Zaczynska, Peter Bourgonje, and Manfred Stede. 2024. How diplomats dispute: The un security council conflict corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8173–8183.
- H Zhang. 2004. The optimality of naive bayes. In *Proceedings of the the 17th International FLAIRS conference (FLAIRS2004)*, pages 562–567.
- Andrea Zielinski, Calvin Spolwind, Henning Kroll, and Anna Grimm. 2023. A dataset for explainable sentiment analysis in the german automotive industry. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 138–148.
- Dan Zigmund. 2021. Toward a computational analysis of the pali canon. *Journal of the Oxford Centre for Buddhist Studies*, 20.

A Training hyperparameters

A.1 Hyperparameters for Devanagari Script

Model	Hyperparameters	Values
Classic Machine Learning Models		
MultinomialNB	Vectorizer: CountVectorizer	binary = False tokenizer = lambda x: x.split() token_pattern = None
LogisticRegression	random_state	0
RandomForestClassifier	random_state	0
SVC (Support Vector Classifier)	probability random_state	True 0
GradientBoostingClassifier	random_state	0
KNeighborsClassifier	n_neighbors	3
TfidfVectorizer Parameters		
All models using TfidfVectorizer	use_idf binary tokenizer token_pattern	True False lambda x: x.split() None
Transformer-Based Models		
XLM-Roberta (plain and fine-tuned)	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps learning_rate warmup_steps weight_decay seed	10 16 steps 100 100 2e-5 500 0.01 0
Electra	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps seed	20 16 steps 100 100 0
T5 (T5-base)	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps learning_rate warmup_steps weight_decay seed	10 16 steps 50 50 2e-5 500 0.01 0
byT5 (byT5-base)	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps learning_rate warmup_steps weight_decay seed	5 8 steps 50 10 2e-5 50 0.01 0
OpenHathi (QLoRA, Sequence Classification)	lora_r lora_alpha lora_dropout bias max_length per_device_train_batch_size gradient_accumulation_steps warmup_steps max_steps learning_rate fp16	128 256 0.1 none 512 8 4 100 2000 4e-5 True

A.2 Hyperparameters for Roman Script

Model	Hyperparameters	Values
Classic Machine Learning Models		
MultinomialNB	Vectorizer: CountVectorizer	binary = False tokenizer = lambda x: x.split() token_pattern = None
LogisticRegression	random_state	0
RandomForestClassifier	random_state	0
SVC (Support Vector Classifier)	probability random_state	True 0
GradientBoostingClassifier	random_state	0
KNeighborsClassifier	n_neighbors	3
TfidfVectorizer Parameters (used in some classic models)		
All models using TfidfVectorizer	use_idf binary tokenizer token_pattern	True False lambda x: x.split() None
Transformer-Based Models		
XLM-Roberta (plain and fine-tuned)	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps learning_rate warmup_steps weight_decay seed	10 16 steps 100 100 2e-5 500 0.01 0
Electra	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps seed	20 16 steps 100 100 0
T5 (T5-base)	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps learning_rate warmup_steps weight_decay seed	10 16 steps 50 50 2e-5 500 0.01 0
byT5	num_train_epochs per_device_train_batch_size evaluation_strategy save_steps logging_steps learning_rate warmup_steps weight_decay seed	5 8 steps 50 10 2e-5 50 0.01 0

Intersecting Register and Genre: Understanding the Contents of Web-Crawled Corpora

Amanda Myntti¹, Liina Repo¹, Elian Freyermuth², Antti Kanner¹,
Veronika Laippala¹, Erik Henriksson¹,

¹University of Turku, ²National Graduate School of Engineering of Caen

Correspondence: amanda.a.myntti@utu.fi

Abstract

Web-scale corpora present valuable research opportunities but often lack detailed metadata, making them challenging to use in linguistics and social sciences. This study tackles this problem by exploring automatic methods to classify web corpora into specific categories, focusing on text registers such as *Interactive Discussion* and literary genres such as *Politics and Social Sciences*. We train two machine learning models to classify documents from the large web-crawled OSCAR dataset: a register classifier using the multilingual, manually annotated CORE corpus, and a genre classifier using a dataset based on Kindle US&UK. Fine-tuned from XLM-R Large, the register and genre classifiers achieved F1-scores of 0.74 and 0.70, respectively. Our analysis includes evaluating the distribution of the predicted text classes and examining the intersection of genre-register pairs using topic modelling. The results show expected combinations between certain registers and genres, such as the *Lyrical* register often aligning with the *Literature & Fiction* genre. However, most registers, such as *Interactive Discussion*, are divided across multiple genres, like *Engineering & Transportation* and *Politics & Social Sciences*, depending on the discussion topic. This enriched metadata provides valuable insights and supports new ways of studying digital cultural heritage.

1 Introduction

Automatically collected web-scale corpora, encompassing billions of words, offer significant opportunities for research across a range of disciplines, including computational linguistics, natural language processing, social sciences, and digital humanities. These extensive resources have been, and continue to be, instrumental in advancing large language models, such as the one underpinning ChatGPT. Additionally, these corpora contain vast amounts of text produced in varied contexts and for diverse

purposes, serving as repositories for new and evolving forms of digital cultural heritage. Consequently, web-scale corpora not only open new research avenues in the humanities and social sciences but also necessitate multidisciplinary collaboration to ensure their effective use (Laippala et al., 2021b; Välimäki and Aali, 2022).

A notable challenge in utilizing web-scale corpora is the lack of detailed metadata describing their contents. Without such metadata, texts of diverse varieties—such as legal notices, advertisements, news articles, fiction, and song lyrics—are treated equally, despite their distinct linguistic characteristics. This complicates the filtering and selection of data relevant to specific research tasks. Among others, these procedures are often crucial for building accurate language models, where the absence of metadata increases the risk of models learning from biased, toxic, or irrelevant data (e.g. Gehman et al., 2020; Carlini et al., 2021; Dodge et al., 2021; Feng et al., 2023; Bannihatti Kumar et al., 2023; Mallen et al., 2023). Text type metadata has also been shown to enhance the performance of various Natural Language Processing (NLP) applications, including part-of-speech taggers, parsers, and information retrieval systems (e.g. Karlgren and Cutting, 1994; Vidulin, 2007; Giesbrecht, 2009; Santini et al., 2011; Van Der Wees et al., 2018; Argamon, 2019).

To address this challenge, we explore a dual approach to classifying web corpora into specific text categories, focusing on two key approaches: register and genre. Registers, as they are typically applied in corpus linguistics, refer to culturally recognizable text varieties characterized by their communicative situation and functionally related linguistic features (Biber 1988; Egbert and Biber 2019; Biber and Egbert 2023). Genres, in literary studies, are often utilized to examine various forms of literary work, focusing on content, context, and narrative tools (e.g., Goyal and Vuppuluri 2022;

Zhang et al. 2022).

In recent years, text classification and specifically web register identification has taken leaps forward, with web register classifiers achieving nearly human-level performance (Laippala et al. 2023; Kuzman et al. 2023b; Henriksson et al. 2024). These advancements enable us to enhance document metadata substantially. However, when register classification is applied to web-scale corpora, the predicted register classes are still very broad and include a wide range of linguistic variation.

Therefore, in this study, we examine whether combining two approaches, namely registers and genres, can enhance the depth of the available information for a given document. Specifically, we examine the intersections between register and genre labels in a text classification setting and explore how these intersections, along with the resulting new metadata, can allow for novel uses of the corpus in other studies. To achieve this, we apply machine learning to train two text classifiers: one targeting registers and the other focusing on genres. These classifiers are then used to predict classes for one million documents from the widely used web-scale OSCAR dataset (Open Super-large Crawled ALMANaCH coRpus; Ortiz Suárez et al. 2019; Laippala et al. 2022).

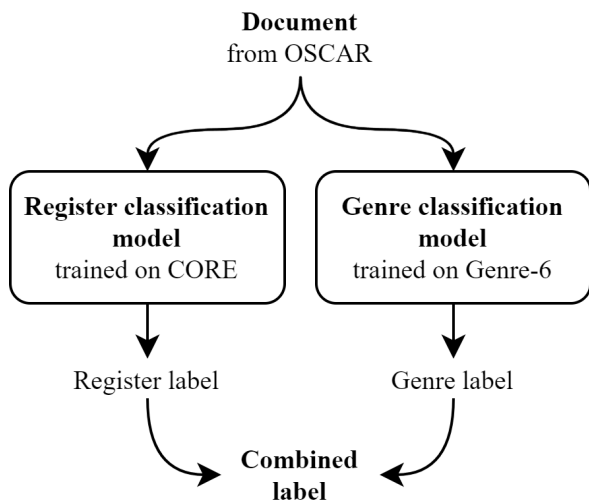


Figure 1: Workflow of our process.

To assess the quality of the new metadata-combined label of register and genre classifiers—we evaluate two conditions. First, we examine the overlap between the two labelling schemes, as cross-labelling has little value if the classifiers produce similar results. Ideally, each register class should map onto multiple genre categories, adding previ-

ously unattainable information to each document. Second, we evaluate whether these cross-labellings are meaningful by using topic modelling to extract topic words for each register-genre combination.

2 Data

We use three different datasets for our experiments, namely the Corpus of Online Registers of English (CORE), the Genre-6 literary genre corpus, and the Open Super-large Crawled ALMANaCH coRpus (OSCAR), each for a different task.

2.1 Register Data

The register classifier is trained using the CORE corpus¹ (Egbert et al. 2015; Laippala et al. 2023), which consists of manually register-annotated English web texts. The corpus contains nearly 50,000 documents and covers the full range of English web registers. The annotation process involved four individual annotators independently assigning each document a main register label and, when possible, a subregister label for a more detailed description. In cases of annotator disagreement, a document could be assigned multiple register labels. The annotation process resulted in a hierarchical multilabel register scheme with eight main register categories with broad, functional labels such as *Narrative*, *Informational description* and *Opinion*, and tens of more detailed subcategories such as *News report*, *Research article* and *Review*. Following Laippala et al. (2022), we slightly modify this hierarchy by mapping some subregisters together to enhance classifier performance. All the main registers and subregister categories of CORE used in this study are shown in Table 1, and the mapping from original CORE labels to our modified labels can be found in Appendix A.

2.2 Genre Data

For the genre classifier, we train using the Genre-6 dataset², which is derived from Kindle UK & US. Genre-6 comprises over 20,000 short stories and novels, with genre categories assigned by the authors. The genre labels are selected from the available categories on Kindle UK & US, resulting

¹Specific version available at <https://github.com/TurkuNLP/CORE-corpus>.

²The original dataset is available at <https://huggingface.co/datasets/marianna13/the-eye>, and the cleaned version used in training is available on our Huggingface page <https://huggingface.co/datasets/TurkuNLP/genre-6>.

Main register	Sub register	Support
How-to/Instruction (HI)		2047
	Recipe (re)	189
	(no subregister)	1858
Interactive discussion (ID)	–	3449
Informational description (IN)		13047
	Description of a thing or a person (dtp)	5444
	Encyclopedia article (en)	556
	FAQ about information (fi)	337
	Legal terms and conditions (lt)	202
	Research article (ra)	936
	(no subregister)	5572
Informational persuasion (IP)		2011
	Description with intent to sell (ds)	1422
	Editorial (ed)	94
	(no subregister)	495
Lyrical (LY)	–	680
Narrative (NA)		21534
	News report (ne)	11785
	Narrative blog (nb)	3620
	Sports report (sr)	3164
	(no subregister)	2965
Opinion (OP)		10754
	Advice(av)	1161
	Opinion blog (ob)	5242
	Reviews (rv)	2065
	Religious blogs/sermons (rs)	776
	(no subregister)	1510
Spoken (SP)		736
	Interview (it)	537
	(no subregister)	199
Total		54258

Table 1: Main and subregister categories of CORE used in this study. Original scheme in [Egbert et al. \(2015\)](#) and mapping to this scheme in [Appendix A](#).

Genre	N
Cookbooks, Food & Wine (Cook)	370
Engineering & Transportation (Engn)	1688
Literature & Fiction (Lit)	4969
Medicine & Health Sciences (Med)	763
Politics & Social Sciences (Pol)	2134
Science & Math (Sci)	1474
None (in prediction only)	–
Total	11398

Table 2: Chosen genre labels of Genre-6 dataset.

in a multilabel genre annotation scheme with categories such as *Children’s Books*, *Science & Math*, and *Action & Adventure*. Initially, we performed minor preprocessing steps to improve data quality, such as excluding categories with minimal support.

Because some of the genre classes in the Genre-6 dataset are overlapping, for this experiment we further chose a subset of genres that maximizes the performance in two ways: Firstly, the chosen genres need to be present in our target corpus, OSCAR. As OSCAR is a web corpus, genres most suitable for our task include common topics in on-

line sources. Secondly, to try to maximize classifier performance, we chose categories by training the classifier with different candidate subsets and evaluating the classifiers’ performance. [Table 2](#) presents the final genre categories used in this study. Lastly, we included a “None” category for uncertain classifications and to address those that fall outside our chosen category scheme, acknowledging that our categories do not fully represent the entire web and some common internet topics, such as religion, are not well covered by this set of labels.

2.3 The Labelled Target Corpus

We use the OSCAR corpus ([Ortiz Suárez et al., 2019](#)) for our analysis of the register–genre intersection. The OSCAR corpus was created by extracting and filtering text from Common Crawl³. It is a multilingual corpus comprising billions of words across 166 languages. For our study, we specifically use the pre-classified version, Register OSCAR⁴ ([Laippala et al., 2022](#)), which has un-

³<https://commoncrawl.org>

⁴https://huggingface.co/datasets/TurkuNLP/register_oscar

dergone further processing, including additional boilerplate text removal, resulting in higher quality than the original OSCAR corpus. We do not use the existing register labels as they only cover the main register level labels, but re-predict the dataset with our register classifier to also get access to the supplementary information of the sublabels. However, in some of our analyses, such as topic modelling, we present results at the level of the main labels for simplicity; in these cases, the sublabels have been aggregated into the main labels according to the label hierarchy shown in Table 1. We label a sample of 1 million documents from OSCAR in this study.

3 Experimental Setup

Figure 1 illustrates the workflow of the classification process. We use two classifiers to predict labels for each document from the OSCAR corpus, and the resulting labels are then combined into a single, combined label.

3.1 Classifier training

We approach classification using a multilabel setting, which previous research on register identification has shown to produce significantly higher scores compared to a single-label, multi-class approach (e.g. Egbert et al. 2015; Madjarov et al. 2019; Sharoff 2021; Laippala et al. 2023). The data is split into training, development, and test sets using stratified sampling, with proportions of 80%, 10%, and 10% for the Genre-6 corpus, and 70%, 10%, and 20% for the CORE corpus. We limit each document to the first 512 tokens, based on (1) the maximum token limit of the models used and (2) the findings of Laippala et al. (2023), which indicate that the best performance in register classification is achieved using the initial 512-token chunks of text.

The register classifier is trained using the CORE corpus, whereas the genre classifier is trained using the Genre-6 corpus, as mentioned in the previous section. Both classifiers are implemented by fine-tuning XLM-RoBERTa-Large (XLM-R; Conneau et al. 2020) for a multilabel classification task using the Huggingface Transformers library. We selected XLM-R due to its efficiency and strong performance in earlier studies of register classification (e.g., Repo et al. 2021). Both models use a multilabel setup and Binary Cross-Entropy Loss, and the prediction threshold is optimized for the F1-

score. We also experimented with Focal Loss (Lin et al., 2018), but ultimately chose Binary Cross-Entropy Loss, as it produced higher-quality predictions in manual evaluation, despite Focal Loss yielding slightly better F1-scores.

3.2 Topic modelling

For our topic modelling experiments, we utilized the Latent Dirichlet Allocation (LDA) algorithm implemented in the `gensim` library⁵. We extracted topics separately from each register–genre intersection class. We also experimented with transformer based BERTopic (Grootendorst, 2022). Using BERTopic, we were able to extract more detailed topics; however, the initial experiments showed that the support of each intersection class influenced the quality of the results, whereas this effect was diminished with the LDA model. Additionally, we preferred the simpler presentation of the LDA model. We used the following parameters: 30 passes, 1 topic, 10 best words per topic. For the largest combination classes, where support is in the hundreds of thousands, we randomly select a subset of 10000 documents for the analysis. We lemmatize, and remove punctuation and English stop words using the `nltk` library⁶. We also experimented with extracting multiple topics for each register–genre combination, which revealed the structure of some intersection classes better. However, for simplicity, we present one topic per class. Additionally, in this part of the experiment, we focus on the main hierarchy level of the register labels to maintain clarity in the presentation.

4 Results

4.1 Classifier Evaluation

Our results show that the register classifier is able to reach an F1-score of 0.74, whereas the genre classifier’s F1-score is 0.70. The class-specific performance of each model is detailed in Tables 3 and 4. Although these scores vary considerably, ranging from 0.45 for *Science & Math* to 0.89 for *Lyrical*, they are consistent with previous register identification results (Egbert et al., 2015), indicating that the predictions are reasonably reliable.

The variability in identification performance can be attributed to how well the registers and genres are defined linguistically, which affects how ac-

⁵<https://radimrehurek.com/gensim/models/ldamodel.html>

⁶<https://www.nltk.org/>

Label	F1-Score	Support
LY	0.8949	135
SP	0.7032	146
ID	0.8475	686
NA	0.8405	4264
HI	0.6788	411
IN	0.7176	2596
OP	0.6854	2129
IP	0.5591	402
it	0.7045	104
ne	0.8120	2359
sr	0.8942	635
nb	0.6745	722
re	0.8116	37
en	0.8079	108
ra	0.6686	189
dtp	0.5271	1090
fi	0.5000	69
lt	0.5763	40
rv	0.7040	411
ob	0.5591	1051
rs	0.7278	157
av	0.5119	236
ds	0.6427	280
ed	0.0000	19
μ (micro)	0.74	18276

Table 3: Results of our trained register classifier. The threshold for classification is set at 0.4, optimized wrt. F1-score.

curately they can be classified (Biber and Egbert, 2018; Biber et al., 2020; Laippala et al., 2021a). For example, *Lyrical* texts, which mostly consist of song lyrics and poems, have distinct characteristics that make them easier to classify accurately. In contrast, texts within the *Advice* subregister vary widely and can be mistaken for other opinionated registers, such as *Opinion Blog*, leading to lower identification scores. A similar pattern is observed in the genre-specific performance, where some classes, like *Science & Math* and *Medicine & Health Sciences*, contain texts that are hard to distinguish clearly between these categories.

4.2 Register and Genre Intersection

Figure 2 illustrates the intersections between the registers and genres. The figure confirms that no register and genre categories fully overlap, demonstrating that cross-labelling with our setup achieves the intended outcome: it refines the classification and enriches the information for each document.

Label	F1	N
Cookbooks, Food & Wine	0.59	35
Engineering & Transportation	0.65	172
Literature & Fiction	0.81	535
Medicine & Health Sciences	0.61	72
Politics & Social Sciences	0.53	194
Science & Math	0.45	144
μ (micro)	0.70	1152

Table 4: Results of our trained genre classifier. The threshold for classification is set at 0.3, optimized wrt. F1-score.

To evaluate the increase in information quantitatively, we calculate mutual information (MI) between the register and genre labels. Mutual information measures the information one label provides about the other and is calculated from the joint probability distribution of the genre and register labels. We use the `scikit-learn`⁷ library to calculate this value, with multilables separated for this step and main-subregister combinations treated as separate classes. Although values of mutual information are not comparable, values close to zero indicate low levels of dependency between the variables, which is ideal in our case. We also calculate the increase of information using Shannon’s entropy H as $\sum_{x \in \mathcal{X}} -P(x) \log_2 P(x)$, which measures the informational value of the variable, with P standing for the marginal probability function is separate label cases and joint probability function in the combined label case. The results are presented in Table 5. These values show that the genre label cannot be inferred from the register label and that the information content increases with the combined labelling compared to using each label scheme separately. Specifically, the additional information contributed by the genre labels to the register labels is $H(\text{genre}|\text{register}) = H(\text{register,genre}) - H(\text{register}) = 2.073$ bits.

MI	$H(\text{register})$	$H(\text{genre})$	$H(\text{register, genre})$
0.109	3.370	2.229	5.443

Table 5: Mutual information (MI) between register and genre and the entropy H of register labels (main and subregisters), genre labels, and the combined labelling.

From Figure 2, expected combinations between certain registers and genres can be seen. For in-

⁷https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif

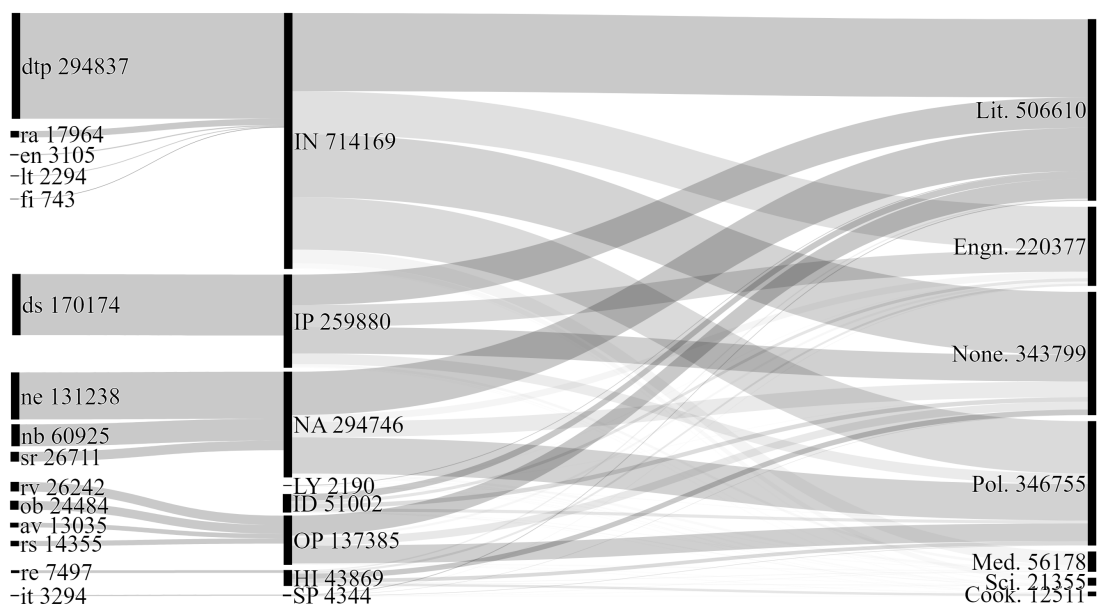


Figure 2: Intersection of registers and genres in the OSCAR corpus. Register sublabels on the left, main level labels on the centre and genres on the right. See Table 1 for register abbreviations and Table 2 for genre abbreviations. The thickness of the connection reflects the relative frequency of co-occurrence, with opacity adjusted to make register-class-wise frequencies more visible. Numbers indicate the number of documents in OSCAR labelled as each class, with multilabels separated. A small number of instances with incorrect register label hierarchy removed.

stance, as could be anticipated, the *Lyrical* register and the *Literature & Fiction* genre co-occur very often, although *Lyrical* also intersects with other genres – such as *Politics & Social Sciences*. Manual evaluation shows this particular intersection class contains religious poetry and lyrical texts with social commentary aspects. An example of this can be seen as the fourth example in Table 7.

Similarly, the *Spoken* register seems to be mostly associated with the *Politics & Social Sciences* and *Literature & Fiction* genres. These combinations convincingly suggest documents with spoken elements, such as conversations or interviews. In *Literature & Fiction*, these might include dialogues between characters, and in *Politics & Social Sciences*, they could be interviews or political speeches. It is also noteworthy that the *Literature & Fiction* class is very large, both in terms of support and the content used for classifier training, and thus covers a variety of texts, including low-quality content. As stated previously, our genre selection criteria were influenced by both the predicted contents of the target documents and the measured classifier performance, which in this case resulted in the selection of the large and sometimes vague class of *Literature & Fiction*.

All other registers are divided into multiple

genres—in particular, the largest registers cover documents across all of them. The intersections are logical; for instance, the largest intersections of *Informational persuasion* are with *Engineering & Transportation* and *Literature & Fiction*, suggesting that these are persuasive documents such as editorials or company websites, discussing topics relevant to these genres, like advertisements for technological devices. Similarly, in addition to these two genres, the largest register class *Informational description* intersects mostly with *Politics & Social Sciences* and *None*, which facilitates the identification of different types of informative/descriptive documents, which were previously only marked by their register characteristics. *Interactive Discussion* intersects with all genres, with the strongest connections to *Engineering & Transportation* and *Politics & Social Sciences*, depending on the discussion topic. Finally, the *Narrative* register also intersects with all these genres. This register primarily includes news articles and narrative blogs, suggesting that these subregisters discuss topics related to these genres. Previously, identifying such combinations required manual searches within the documents, but with the addition of genre labels, these intersections can now be easily categorized, filtered, or selected from the corpus.

From the perspective of the genres, similar results can be seen. For example, as expected, the genre *Medicine & Health Sciences* mostly intersects with *Informative description*, however, we were able to extract conversations and ads with health care related topics with the intersection of registers *Interactive discussion* and *Informational persuasion*.

Finally, the *None* category for uncertain genre predictions contains documents from all the register classes, most notably from *Informational description* and *Informational persuasion*. This is expected due to the class size imbalance, however, these classes are over-represented in the *None* genre category compared to the *Narrative* register class, which despite its large size rarely intersects with the *None* genre. The reasons behind this lie in the variety of topics discussed in the respective registers. The registers frequently intersecting with the *None* genre seem to cover topics not included in the genres, or the documents within the register simply do not feature a well-defined genre at all. For instance, in Table 7, a text combining *Interactive discussion* with *None* a disproved message on a forum, which does not fit any of the specified genres. This shows that the *None* class is linguistically motivated and provides meaningful intersections.

4.3 Contents of the intersection classes

We use topic modelling to gain insights into each register–genre intersection class, and the results are presented in Table 6. Some register–genre combination classes produced topic keywords that reflect features of both the register and the genre involved. For instance, most keywords extracted from combinations with the register *How-to Instructions* include characteristic verbs such as “make” and “use”, which are frequently used in instructive texts. Similarly, the *Interactive discussion* register produced topic words containing personal pronouns like “I(’m)” and “us”, which are typical of discussions. In the *Informational persuasion* register, words like “product”, “help”, and “business” suggest that the texts feature ads and other commerce-related documents. The *Informational description* register is associated with verbs in the past tense, which is characteristic of texts such as encyclopedic entries.

From the perspective of the genre classes, similarities are shared over register class boundaries. For instance, “food”, a characteristic word for the *Cookbooks, Food & Wine* genre, appears in all but

Reg.	Genre	N	Topic keywords
HI	Cook.	6838	make, add, recipe, minutes, use
	Engn.	5875	use, one, need, make, new
	Lit.	3815	like, also, love, new, would
	Med.	877	skin, may, help, make, like
	Pol.	10219	use, also, make, time, may, new
ID	Sci.	975	time, water, may, make, need
	Cook.	256	rustic, home, plans, house, would
	Engn.	8759	would, get, new, time, need, work
	Lit.	18947	would, i’m, know, good, think
	Med.	1278	would, time, also, good, i’m
IN	Pol.	8300	time, people, need, us, know
	Sci.	1001	one, would, know, new, good
	Cook.	2228	food, used, make, made, many
	Engn.	99490	new, time, used, us, system, data
	Lit.	190734	dating, free, pdf, online, first
IP	Med.	35252	patients, health, care, treatment, new
	Pol.	134903	people, use, get, us, information
	Sci.	12814	one, two, water, used, species
	Cook.	1396	food, like, product, used, new
LY	Engn.	41796	new, us, service, quality, system
	Lit.	62779	free, get, great, book, home
	Med.	4352	skin, health, help, body, new
	Pol.	21401	new, business, us, help, people
NA	Sci.	1112	fishing, new, plants, water, use
	Lit.	1908	download, love, would, know, i’m
	Cook.	427	make, food, good, would, made
	Engn.	17251	said, would, time, first, us
	Lit.	113824	new, would, first, back, people
OP	Med.	5983	new, health, said, people, may
	Pol.	93575	said, would, people, us, first
	Sci.	3602	new, said, would, water, two
	Cook.	705	food, make, good, get, great,
	Engn.	5652	new, time, car, make, well
SP	Lit.	52402	would, also, us, people, first
	Med.	2697	get, people, would, may, help
	Pol.	45086	people, would, time, us, new
	Sci.	274	one, people, new, us, many
SP	Engn.	147	would, think, work, time. people
	Lit.	2060	think, people, would, really, know
	Pol.	1370	people, think, know, us, going

Table 6: Selected 5 of 10 top topic words extracted from the intersection classes. “None” class and classes with less than 100 documents omitted. See Table 1 for register abbreviations and Table 2 for genre abbreviations.

one of the combination classes. Interestingly, the intersection of *Cookbooks, Food & Wine* and *Interactive discussion* contains more documents seemingly about kitchen decor, not only about food. In the *Medicine & Health Sciences* genre, themes present throughout are patients and helping, while in *Science & Math*, recurrent topic words include

Register	Genre	Truncated document
IN, ra	Engineering & Transportation	The management of existing road infrastructures is a multidisciplinary activity that involves structural engineering, material science, management, economics and ecology. The objective is to achieve maximum availability of road links at minimum societal costs. Recently, tools (Bridge Management Systems, BMSs) have been developed to help decision makers to determine the optimal management strategies within available resources.
NA, ne, IN, dtp	Politics & Social Sciences	Second-year medical student Walter Humann is the winner of the 2015 Danny Jones History of the Health Sciences Student Essay Competition. A panel of three judges rated the six entries on the quality of writing, the comprehension of the issues, the clarity of discussion, and the applicability of the topic to the history of health care. His essay Medical Progress in the West: A Historical Perspective addresses the three eras of history in which medical developments struggled against societal norms.
OP	Cookbooks, Food & Wine	White chocolate isn't really chocolate at all. While it contains the cocoa butter of true chocolate, it lacks cocoa solids, the element responsible for milk and dark chocolate's characteristic brown color and nutty roasted flavor. Other pale confections labeled simply "white" chips or bars (these boast less than the 20 percent cocoa butter required to earn the designation "white chocolate") are just as common in the baking aisle of the supermarket.
LY	Politics & Social Sciences	I am obnoxious to each carping tongue Who says my hand a needle better fits. A Poet's Pen all scorn I should thus wrong, For such despite they cast on female wits. If what I do prove well, it won't advance, They'll say it's stol'n, or else it was by chance.
ID	None	I posted a question to the forum an hour ago and received an email saying the moderators had approved the content. Checking a few minutes ago I see a notice that the message has now be disproved. How do I contact a moderator to find out what is going on?

Table 7: Examples of our classification results. Texts truncated and original spelling retained. See register abbreviations in Table 1.

“water” and “plants”.

Some drawbacks can also be seen from these topic keywords. The keywords for the genre classes *Literature & Fiction* and *Politics & Social Sciences* offer few distinguishing words, apart from references to people. Manual evaluation confirms this observation; as previously noted, the *Literature & Fiction* class, in particular, contains a large variety of texts, and thus extracting a single topic from this class does not result in coherent keywords. In the case of *Lyrical* and *Spoken* registers, the small support affects the results of the topic modelling. However, for the *Lyrical* register, keywords like “love” and “I’m” are characteristic of song lyrics and poetry; the keyword “download” likely appears due to boilerplate text commonly found on song lyrics web pages.

Table 7 presents example documents associated with the register–genre intersections. The first example, labelled as a *Informational description – Research article* and *Engineering & Transportation* reflects the extracted topics, including topic keywords like “system”. In the second example, labelled as *Informational description – Description of a thing or a person, Narrative – News report* for register and *Politics & Social Sciences* for genre, clearly contains characteristics of a news article and describes a person. At first glance, the text could be labelled as *Medicine & Health Sciences*, however,

the document actually covers history and struggles against societal norms. In the third example, the vocabulary clearly reflects the class *Cookbooks, Food & Wine*, but the tone is correctly identified as opinionated, as the text contains emotionally charged adjectives, which are also seen in the topic words extracted for the register class *Opinion*. This example also justifies our decision to include the *Cookbooks, Food & Wine* genre, despite the CORE register scheme already containing a *Recipe* sub-register, as it allows us to capture a broader range of food-related documents.

5 Conclusion

In this study, we experimented with labelling a large internet corpus using two classifiers and evaluated the new metadata produced by the intersection of two classification schemes. We trained the classifiers on available register and genre datasets in a multilabel classification setting and reached reliable results. We then analyzed the distributions of the intersection classes and extracted topic keywords from them. Our evaluation, based on quantitative analysis using topic modelling and close reading, demonstrated that the predicted genre and register labels provide meaningful auxiliary information, facilitating new ways to use the corpus.

This is particularly valuable for digital humanities and cultural heritage studies, as it allows for

richer contextualization and more nuanced analysis of historical documents, literary texts, and other cultural artefacts. Enhanced metadata can also support the preservation and accessibility of digital archives, ensuring that documents are more easily discoverable and interpretable.

In the future, we will aim to improve model performance by experimenting with different model architectures and refining the chosen classes of the genre classifier. Specifically, we recognize that the current *Literature & Fiction* category is too broad and plan to subdivide it into more specific genres. We are also interested in exploring different data augmentation techniques (e.g. label cleaning tools⁸), particularly for our genre corpus, which has shown apparent label issues during manual review.

6 Limitations

Our work is conducted entirely in English. While previous studies (e.g. Repo et al., 2021; Rönnqvist et al., 2021) suggest that the effectiveness of register classification using the CORE scheme may transcend language barriers, the same may not apply to our genre classification system. Additionally, we based our genre classification training on methods typically used for registers, which may not perfectly align with genre distinctions. Our selection of genre categories relied on the support of the classes, partly due to the limited number of instances in the corpus. As previously mentioned, we recognize the bias towards more technical genres, as the selected genre categories contain both engineering and science related topics but lack coverage of other common internet subjects. Another possible approach to text classification would have been to use recent large language models such as ChatGPT⁹. Kuzman et al. (2023a) compared the performance of an XLM-R-based model to GPT-3.5 and GPT-4 (OpenAI, 2023) in register (genre in their terminology) classification. In their experiments, GPT-4 and XLM-R-Large performed similarly on out-of-domain English testset. This indicates that using GPT-like models for this task holds substantial potential.

Acknowledgements

We wish to acknowledge FIN-CLARIAH (Common Language Resources and Technology Infrastructure), and CSC – IT Center for Science for

computational resources. This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Academy of Finland under grant numbers 358720 and 331297. We thank LAION¹⁰ and Ontocord.AI¹¹ for collaboration on the genre dataset.

References

- Shlomo Engelson Argamon. 2019. *Computational register analysis and synthesis*. *ArXiv*, abs/1901.02543.
- Vinayshekhar Bannihatti Kumar, Rashmi Gangadharahiah, and Dan Roth. 2023. *Privacy adhering machine un-learning in NLP*. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 268–277, Nusa Dua, Bali. Association for Computational Linguistics.
- Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.
- Douglas Biber and Jesse Egbert. 2018. *Register Variation Online*. Cambridge University Press.
- Douglas Biber and Jesse Egbert. 2023. *What is a register?: Accounting for linguistic and situational variation within – and outside of – textual varieties*. *Register Studies*, 5.
- Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. *Reconceptualizing register in a continuous situational space*. *Corpus Linguistics and Linguistic Theory*, 16(3):581–616.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. *Extracting training data from large language models*. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jesse Dodge, Ana Marasovic, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. *Documenting large webtext corpora: A case study on the colossal clean crawled corpus*. In *Conference on Empirical Methods in Natural Language Processing*.

⁸<https://github.com/cleanlab/cleanlab>

⁹<https://openai.com/chatgpt/>

¹⁰<https://laion.ai/>

¹¹<https://www.ontocord.ai/>

- Jesse Egbert and Doug Biber. 2019. [Incorporating text dispersion into keyword analyses](#). *Corpora*, 14(1):77–104.
- Jesse Egbert, Douglas Biber, and Mark Davies. 2015. [Developing a bottom-up, user-based method of web register classification](#). *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. [From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.
- Eugenie Giesbrecht. 2009. In search of semantic compositionality in vector spaces. In *Conceptual Structures: Leveraging Semantic Technologies*, pages 173–184, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Anshaj Goyal and Prem Vuppuluri. 2022. [Statistical and Deep Learning Approaches for Literary Genre Classification](#), pages 297–305. Springer Singapore, Singapore.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Erik Henriksson, Amanda Myntti, Anni Eskelinen, Selen Erten-Johansson, Saara Hellström, and Veronika Laippala. 2024. [Untangling the unrestricted web: Automatic identification of multilingual registers](#).
- Jussi Karlgren and Douglass Cutting. 1994. [Recognizing text genres with simple metrics using discriminant analysis](#). In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.
- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023a. [Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models](#). *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.
- Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2023b. [Get to know your parallel data: Performing English variety and genre classification over MaCoCu corpora](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103, Dubrovnik, Croatia. Association for Computational Linguistics.
- Veronika Laippala, Jesse Egbert, Douglas Biber, et al. 2021a. [Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents](#). *Language Resources and Evaluation*, 55:757–788.
- Veronika Laippala, Aki-Juhani Kyröläinen, Jenna Kanerva, and Filip Ginter. 2021b. [Dependency profiles in the large-scale analysis of discourse connectives](#). *Corpus Linguistics and Linguistic Theory*, 17(1):143–175.
- Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo. 2023. [Register identification from the unrestricted open web using the corpus of online registers of english](#). *Language Resources and Evaluation*, 57(3):1045–1079.
- Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika, and Sampo Pyysalo. 2022. [Towards better structured and less noisy web data: Oscar with register annotations](#). In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. [Focal loss for dense object detection](#). *Preprint*, arXiv:1708.02002.
- Gjorgji Madjarov, Vedrana Vidulin, Ivica Dimitrovski, and Dragi Kocev. 2019. [Web genre classification with methods for structured output prediction](#). *Information Sciences*, 503:551–573.
- Alex Mullen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khachabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4](#).
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. [Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures](#). In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019, Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. [Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for*

Computational Linguistics: Student Research Workshop, pages 183–191, Online. Association for Computational Linguistics.

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. [Multilingual and zero-shot is closing in on monolingual web register classification](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Marina Santini, Alexander Mehler, and Serge Sharoff. 2011. [Riding the Rough Waves of Genre on the Web](#). In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 3–30. Springer Netherlands, Dordrecht.

Serge Sharoff. 2021. [Genre annotation for the Web: Text-external and text-internal perspectives](#). *Register Studies*, 3(1):1–32.

Marlies Van Der Wees, Arianna Bisazza, and Christof Monz. 2018. [Evaluation of machine translation performance across multiple genres and languages](#), pages 3822–3827. European Language Resources Association (ELRA). International Conference on Language Resources and Evaluation, LREC ; Conference date: 07-05-2018 Through 12-05-2018.

Vedrana Vidulin. 2007. [Training the genre classifier for automatic classification of web pages](#). *Journal of Computing and Information Technology*, 15.

Reima Välimäki and Heta Aali. 2022. [The Ancient Finnish Kings and their Swedish Archenemy: Nationalism, Conspiracy Theories, and Alt-Right Memes in Finnish Online Medievalism](#), pages 55–78. *Studies in Medievalism XXXI*. Boydell and Brewer, Boydell and Brewer.

Jinbin Zhang, Yann Ciarán Ryan, Iiro Rastas, Filip Ginter, Mikko Tolonen, and Rohit Babbar. 2022. [Detecting sequential genre change in eighteenth-century texts](#). In *Proceedings of the Computational Humanities Research Conference 2022*, CEUR Workshop Proceedings, pages 243–255, Germany. CEUR-WS.org.

A CORE label scheme modification

The mapping from CORE original label scheme to the one used in this study is presented in Table 8.

	Original CORE	Simplified scheme
Register	Subregister	
Narrative (NA)	News report/blog (ne)	–
	Sports report (sr)	–
	Personal/diary blog (pb)	Narrative blog (nb)
	Historical article (ha)	Narrative (NA)
	Travel blog (tb)	Narrative blog (nb)
	Short story (ss)	Narrative (NA)
	Novel	Narrative (NA)
	Biographical story/history	Narrative (NA)
	Magazine article (ma)	Narrative (NA)
	Obituary	Narrative (NA)
	Memoir	Narrative (NA)
	Other narrative (on)	Narrative (NA)
	Opinion (OP)	Opinion blog (ob)
Review (rv)		–
Religious blog/sermon (rs)		–
Advice (av)		–
Letter to the editor (le)		Opinion (OP)
Self-help		Opinion (OP)
Advertisement (ad)		Opinion (OP)
Other opinion (oo)		Opinion (OP)
Informational Description (IN)	Description of a thing (dt)	Description of a thing or a person (dtp)
	Informational blog (ib)	Informational Description (IN)
	Description of a person (dp)	Description of a thing or a person (dtp)
	Research article (ra)	–
	Abstract	Informational Description (IN)
	FAQ about information (fi)	–
	Legal terms and conditions (lt)	–
	Course materials (cm)	Informational Description (IN)
	Encyclopedia article (en)	–
	Technical report (tr)	Informational Description (IN)
	Other informational (oi)	Informational Description (IN)
Interactive Discussion (ID)	Discussion forum (df)	Interactive Discussion (ID)
	Question/answer forum (qa)	Interactive Discussion (ID)
	Reader/viewer responses (rr)	Interactive Discussion (ID)
	Other interactive discussion (of)	Interactive Discussion (ID)
How-to Instructional (HI)	How-to (ht)	How-to Instructional (HI)
	Recipe (re)	–
	Instructions	How-to Instructional (HI)
	FAQ about how-to (fh)	How-to Instructional (HI)
	Technical support (ts)	How-to Instructional (HI)
Other how-to/instructional (oh)	Other how-to/instructional (oh)	How-to Instructional (HI)
	Description with intent to sell (ds)	–
	Persuasive article or essay (pa)	Informative Persuasion (IP)
	Editorial (ed)	–
Other informational persuasion (oe)	Other informational persuasion (oe)	Informative Persuasion (IP)
Lyrical (LY)	Song lyrics (sl)	Lyrical (LY)
	Poem (po)	Lyrical (LY)
	Prayer (pr)	Lyrical (LY)
	Other lyrical (ol)	Lyrical (LY)
Spoken (SP)	Interview (it)	–
	Transcript of video/audio (ta)	Spoken (SP)
	Formal speech (fs)	Spoken (SP)
	TV/movie script (tv)	Spoken (SP)
	Other spoken (os)	Spoken (SP)

Table 8: Mapping from original CORE scheme to the scheme used in this study. Dashes indicate the subregister was preserved identically. Subregisters without abbreviations have zero support in the specific version of CORE that we use.

Sui Generis: Large Language Models for Authorship Attribution and Verification in Latin

Gleb Schmidt
Radboud University
Nijmegen, Netherlands
gleb.schmidt@ru.nl

Svetlana Gorovaia
LEYA Lab, HSE University
St. Petersburg, Russia
sgorovaya@hse.ru

Ivan P. Yamshchikov
CAIRO, THWS
Würzburg, Germany
ivan.yamshchikov@thws.de

Abstract

This paper evaluates the performance of Large Language Models (LLMs) in authorship attribution and authorship verification tasks for Latin texts of the Patristic Era. The study showcases that LLMs can be robust in zero-shot authorship verification even on short texts without sophisticated feature engineering. Yet, the models can also be easily “mislead” by semantics. The experiments also demonstrate that steering the model’s authorship analysis and decision-making is challenging, unlike what is reported in the studies dealing with high-resource modern languages. Although LLMs prove to be able to beat, under certain circumstances, the traditional baselines, obtaining a nuanced and truly explainable decision requires at best a lot of experimentation.

1 Introduction

Unlike in computational linguistics, authorship analysis in the field of digital humanities still largely relies on the complicated process of domain-specific manual feature engineering (Corbara et al., 2020; Manousakis and Stamatatos, 2023; Corbara et al., 2023; Clérice and Glaise, 2023; Beullens et al., 2024). This is mostly due to the fact that the predictions made by machine learning models with regard to philological and historical authorship problems are expected to be contextualized within long-standing scholarly traditions with their established views on what kind of features matter in the studied material (Clérice and Glaise, 2023). For this reason, although deep-learning approaches, especially those based on pre-trained language models (Fabien et al., 2020; Rivera-Soto et al., 2021; Ai et al., 2022; Huertas-Tato et al., 2022; Yamshchikov et al., 2022; Wang et al., 2023; Huertas-Tato et al., 2024), have recently demonstrated their reliability and high performance, even in the most sophisticated settings of virtually all authorship-related

tasks¹ offered at PAN competitions² (Stamatatos et al., 2023; Petropoulos, 2023; Guo et al., 2023).

Pre-trained language models offer valuable insights even in challenging scenarios such as authorship analysis with limited training data, cross-discourse type verification or attribution, style change detection, and cases of stylistic mimicry where authors deliberately disguise their writing style. Additionally, there has been a recent surge in explainable Artificial Intelligence (XAI) techniques, including feature ranking, probing, factual and counterfactual selection, attribution maps, and concept relevance propagation (Achtibat et al., 2023). While these methods are neither flawless (Setzu et al., 2024) nor exhaustive, they represent a significant advancement in the field of explainability.

The linguistic “knowledge” of LLMs, acquired through training on extensive multilingual textual datasets, along with their advanced inference capabilities and their ability to provide human-like natural language explanations for their outputs inevitably raise the question of how these systems can be leveraged for philological and historical investigations.

To promote the wider adoption of large language models (LLMs) as research tools in the digital humanities, this study assesses the zero-shot performance of several publicly available, state-of-the-art LLMs — namely GPT-4o, Gemini, Mistral, and Claude — in authorship verification and attribution tasks. In contrast to previous studies, which have primarily worked with modern languages, our research focuses on a historical language: Latin. To the best of our knowledge, this is one of the first studies to evaluate GPT-4o’s “proficiency” in Latin and the first to report test results for the three

¹As of 2024, the following tasks have been offered at least once: Authorship Attribution, Authorship Clustering, Generative AI Detection, Authorship Verification, Authorship Obfuscation Author Profiling, and Author Diarization.

²<https://pan.webis.de/>

additional models.

Our experiments seek to evaluate the zero-shot effectiveness of LLMs in authorship verification and attribution. We compare their performance against traditional baselines, including classical machine learning classifiers and models based on the pre-trained Latin transformer, LaBerta. Furthermore, we investigate how variations in the quantity and nature of instructions prompted to the LLMs impact the accuracy of their predictions.

2 Related Work

Authorship attribution and authorship verification are two fundamental tasks in authorship analysis. They are the most popular applications of stylometry — the modelling of writing style using statistical methods. Attribution, in its simplest form, aims to identify the author of a previously unseen text sample from a list of candidate authors. Verification, on the other hand, involves determining whether two given texts were written by the same person. Both tasks can vary in complexity, especially when additional challenges arise, such as cross-domain or cross-discourse type problems.

The origins of stylometric analysis for authorship-related problems go back to the 1960s (Mosteller and Wallace, 1963). Early work in stylometry for authorship attribution relied on extraction of hand-crafted features believed to represent the writing style (word frequency, sentence length, and syntactic patterns, etc.) and Bayesian inference (Mosteller and Wallace, 1984; Holmes, 1994, 1998). The study by Joachims (1998) on text categorization provided a foundation for applying Support Vector Machines (SVMs) to stylometric data. In 2002, Burrows suggested a distance-based technique, which became formative for the present-day stylometry, the method now known as Burrows’ Delta. Since then, various features and classification methods were used to quantify stylistic differences and estimating the likelihood of shared authorship between texts. The work of Stamatatos (2009) provides a comprehensive overview of the classical methods used in stylometric analysis.

Since 2010s, the evolution of methodologies for solving these tasks can be traced through the overviews of PAN competitions. Until 2016, with rare exceptions, texts in authorship analysis were treated as bags-of-words (Stamatatos, 2009; Koppel et al., 2009). Research in this field revolved

around various stylistic features such as word and character n-grams, sentence lengths, word and punctuation frequencies, part-of-speech (POS) tag frequencies, and POS n-grams (Stamatatos, 2013). These features were often combined with feature selection or weighting mechanisms and utilized alongside distance measures and standard classifiers like Support Vector Machines (SVM) or Naive Bayes.

The rise of the neural networks marked the shift towards closer attention to the sequential nature of the text. Convolutional and Recurrent Neural Networks (CNNs and RNNs) and later transformers have proven outperform the previous methods, particularly in cases where the writing style is more nuanced and complex (Shrestha et al., 2017; Kestemont et al., 2018, 2020). Yet, this improvement was achieved at expense of the models’s explainability.

Transformer models, such as BERT, RoBERTa, and T5, made authorship attribution and verification systems particularly robust. Since the advent of Siamese network architectures (Reimers and Gurevych, 2019) and the work presented by Fabien et al. (2020) fine-tuning pre-trained models to solve authorship problems has de facto become a standard approach (Rivera-Soto et al., 2021; Stamatatos et al., 2022; Ai et al., 2022; Huertas-Tato et al., 2022; Yamshchikov et al., 2022; Stamatatos et al., 2023; Wang et al., 2023; Huertas-Tato et al., 2024), although ensemble models integrating additional stylometric features (Fabien et al., 2020; Ai et al., 2022) and even independent use of manually engineered features remain quite common (Manousakis and Stamatatos, 2023; Corbara et al., 2023; Clérice and Glaise, 2023; Beullens et al., 2024; Camps et al., 2024).

Since the release of GPT-3.5 in late 2022 (Brown et al., 2020) and the subsequent emergence of GPT-4 (Achiam et al., 2023) and a pleiade of LLMs (Naveed et al., 2023), there have been numerous reports of their groundbreaking performance on various research tasks relevant for the humanities (Karjus, 2023). These tasks range from relatively simple data processing, cleaning, and structuring tasks (such as post-OCR correction, NER, and markup) to data augmentation and labeling (Törnberg, 2024), from semantic search to confirmatory topic analysis (Oiva et al., 2023), and from text summarization and translation (Volk et al., 2024) to multimodal processing. The examples of successful applications continue to proliferate, paving the

way for what [Karjus](#) has described as “machine-assisted mixed methods” (2023), which facilitate interaction with data and promise unprecedented scaling of research efforts.

Ironically, although the availability of LLMs made the detection of machine-generated text one of the most relevant real-world tasks for linguistic forensics and consequently the prevalent topic at PAN competitions ([Bevendorff et al., 2024](#)), the number of studies which explore the LLMs’s own abilities to solve authorship-related problems or serve for feature extraction is rather limited so far.

[Hicke and Mimno \(2023\)](#) leveraged a pre-trained T5 model further fine-tuning for authorship identification in Early Modern English drama. [Patel et al. \(2023\)](#) tried to bridge the gap between stylometry and language models annotating examples of writing style and creating interpretable machine-generated writing style embeddings. A somewhat comparable approach was also proposed in [Ramnath et al. \(2024\)](#). The model is trained using a distillation process from GPT-4-Turbo to Llama-3-8B model. First, GPT is used to produce and standardize a corpus of structured writing style descriptions. Llama is then fine-tuned to produce similar descriptions. This approach addresses the challenges of interpretability in authorship analysis by trying to establish a clear and consistent framework for it.

An immediate source of inspiration for this study, the work by [Huang et al. \(2024\)](#), focuses on a direct prompting of different models with authorship-related questions. The authors arrived at the conclusion that guiding the model by explicitly providing specific linguistic features to pay attention to can significantly improve the precision of the model’s prediction and the quality of the analysis.

We find a compelling reason to explore the use of LLMs to be the challenge posed by sample size. Traditional machine learning methods, such as those described by [Eder \(2015\)](#) and [Eder \(2017\)](#), often require samples of approximately 1000 words to achieve reliable results. Deep learning approaches typically require substantial amounts of training data, which can be difficult to obtain. In contrast, LLMs can perform effectively without extensive additional training, making them advantageous when dealing with limited or costly data resources.

While the impressive results reported in some studies ([Fabien et al., 2020](#); [Kestemont et al., 2019](#))

are noteworthy, it is important to stress that they were conducted using English-language datasets. Given the widespread use of English in the training data of state-of-the-art LLMs, there is a possibility that some of the datasets may overlap with the training data, potentially influencing the outcomes ([Brown et al., 2020](#)).

The case of Latin is very different. First of all, the overall amount of available data is incomparably less. Second, it remains unknown how much of it is actually in the training data of the major LLMs. In 2023, [Burns](#) evaluated the amount of Latin in the training dataset of GPT-3.5 as 339 million tokens, assuming that this number could be higher for GPT-4o. Although Latin is arguably the highest-resourced of all the historical languages, the extent of the easily-available Latin dataset hardly exceeds 700 million words ([Bamman and Burns, 2020](#)) (including Neo-Latin, Latin Wikipedia, and Internet Archive), while state-of-the-art language models for this language are trained on even smaller data, a clean and high-quality subset of the extant corpus, mostly ([Roelli, 2014](#); [Riemenschneider and Frank, 2023a](#); [Ströbel, 2022](#); [Bamman and Burns, 2020](#)). Nevertheless GPT-4 excels in various tasks involving Latin, such as morpho-syntactic annotation (tagging), translation from and into Latin, as well as in text summarization and paraphrasing ([Volk et al., 2024](#)).

To the best of our knowledge, no comprehensive study has yet been conducted on the performance of major LLMs, such as GPT-4, in the specific tasks of authorship verification and attribution in Latin. Similarly, the capabilities of other mature LLMs released after GPT-4, such as Gemini, Claude, and Mistral, have also not been thoroughly examined in this context.

For this reason, in our investigation, we tried not only to measure LLMs’s performance and compare it to conventional baselines but also to study the discrepancy between different LLMs.

3 Methodology

We conducted a series of experiments on two tasks: authorship attribution and authorship verification. The experiments utilized direct prompting of the flagship versions³ of four major LLMs: GPT-4o, Claude, Mistral-Large, and Gemini-1.5. However, only the authorship verification experiments involved all four models, as only GPT-4o demon-

³As of July-August 2024.

strated competitive results in the preliminary authorship attribution tests.

All prompting was implemented in a model-agnostic manner using LangChain library. The choice was mostly dictated by the fact that this library offers a unified API to interact with many different models and facilitates crucial operations such as rate limiting, error handling (request retries), fallbacks, and, most importantly, obtaining structured output from the models.

Each run assessed the performance of a specific model on a given task within a particular setting, defined by the prompt used.

We tested three settings differed by the level of guidance the models received in addition to the default task definition:

1. **BASE**: the models get only a general description of the task;
2. **LIMITED**: the models get a general description of the task and explicit instruction to pay attention to writing style;
3. **HIP**: historically informed prompting, when the models get a general description of the task and a concise list of features to pay attention to formulated by a domain expert and anchored into the scholarly tradition.

Each of the aforementioned settings was tested in two different variants: basic and topic-ignorant, in which the models were explicitly instructed to avoid taking the content and theme into account. For the exact formulation of the prompts, see Appendix B.

To gain further insight into the models’s decision-making processes and compare their performance, we undertook two additional steps: (1) we investigated the influence of semantic similarity on the predictions, and (2) we measured the agreement between the models. For the former, the texts used in our Authorship Verification experiments were vectorized using OpenAI’s text-embedding-3-large model, and a pairwise cosine similarity was calculated between them. We then computed the correlation between these similarity scores and the the models’ predictions across various prompt settings. For the latter, we calculated pairwise joint probability of agreement between models, the pairwise agreement scores are presented in A.

Author	Word count
Augustine of Hippo	2,519,484
Gregory the Great	794,955
Origen (tr. by Rufinus)	385,346
Caesarius of Arles	311,965
Petrus Chrysologus	189,864
Jerome	178,704
Optatus of Milevis	169,280
Quodaultdeus	132,160
Chromatius of Aquileia	115,446
Venerable Bede	114,282
Leo the Great	113,373
Maximus of Turin (I)	73,836
Gregory of Elvira	70,712
Zeno of Verona	48,077
Gaudentius of Brescia	46,127
Ambrose of Milan	43,118
Valerian of Cimiez	31,352
Basil of Caesarea (tr. by Rufinus)	31,240
Priscillian of Avila	23,165
Fulgentius of Ruspa	14,804
Leander of Seville	13,077
Potamius of Lisbon	9,190

Table 1: Dataset.

3.1 Metrics and Baselines

To evaluate the performance of the models, we relied on accuracy, precision, recall, and F1 score.

Furthermore, the performance of the LLMs on each of the two tasks was compared against two different baselines, (four baselines in total). For each task, one baseline features a classical machine learning approach, while the other builds upon a state-of-the-art pre-trained transformer model for Latin, LaBerta (Riemenschneider and Frank, 2023b). For details, see the Tables 3, 7, and 4.

4 Data

In this study, we focus on a subset of the Patristic Sermon Textual Archive (PaSTA), a corpus of Latin homiletic literature of the Patristic era. We prefer this corpus to a seemingly more conventional Classical Roman prose for a reason. Indeed, the very nature of the genre of sermon (or homily) — oral and written — provides a wide spectrum of styles depending on the occasion on which sermons were delivered, the intended audience, underlying material, etc. At the very same time, the act of preaching was always framed by the scriptural and liturgical context. As the goal of the preacher was to ex-

plain the message of the Scripture, demonstrate its relevance to the everyday lives of the flocks, and make clear the symbolic and moral meaning of the sacraments and feasts of the Church, the creativity of the preacher was constantly confronted with the canons of the established genre, which suggested themes as well as discursive and rhetorical devices (Boodts and Schmidt, 2022). Such relative thematic homogeneity of the homiletic corpus makes it a particularly interesting and complex benchmark.

4.1 Preparation and General Preprocessing

For the sake of quality, the data was extracted not only from various open (*Patrologia Latina* as available in the *Corpus Corporum*) resources but also proprietary ones (*Corpus Christianorum Series Latina*), which is why we cannot publish the full texts along with all the associated rich metadata. However, we provide all the data used in the described experiments — the randomly sampled textual fragments with the corresponding author labels. All the data is published on GitHub⁴.

Out of the 62 distinct authors currently represented in PaSTA, we selected 22 authors featured in Dupont et al. (2018), a standard reference work to survey Latin preaching from the 3rd to the 7th centuries, see Table 1. This selection covers all regions of the Late Antique Latin West and encompasses all homiletic subgenres.

Since most of the texts used in the study constitute composite entities (e.g., collections of sermons, epistles, gatherings of treatises, etc.), we first divided all the material into units (henceforth, work-units) representing self-contained acts of preaching (e.g., *sermo*, *homilia*, *tractatus*, epistle, *dictio*). Subsequently, for different experiments, the texts were split into chunks of *approximately* (1) 250 and (2) 500 words. We opted for an oscillating chunk length to respect sentence boundaries. Therefore, some chunks are slightly longer or shorter than the target length.

Sampling texts from the pools of chunked examples was done for each task independently.

4.2 Authorship Verification

Before conducting the first authorship verification experiment, we sampled 5 positive and 5 negative pairs for each of the 22 authors. This process was repeated three times, allowing us to perform each

⁴https://github.com/glsch/sui_generis.

experiment with three distinct sets of pairs. While each pair was unique, the same passage could appear in multiple pairs. This yielded a balanced corpus of 660 pairs, with 30 pairs per author — 15 positive and 15 negative. This same set of 660 pairs⁵ was used across all subsequent authorship verification experiments, with 220 pairs evaluated in each iteration, though the content of each iteration could vary depending on the model employed.

4.3 Authorship Attribution

Authorship attribution experiments were conducted using varying numbers of candidate authors: 5, 10, 15, and 22. For each of these configurations, we randomly selected the required number of authors. To ensure diversity and enhance the reliability of the results, this selection process was repeated five times, generating distinct sets of candidate authors for each iteration.

The sampling of text examples proceeded as follows. For each randomly selected author, we randomly picked two text fragments. The first fragment was designated as the query text, while the second fragment, drawn from a different work by the same author, served as the target text (i.e., the text forming a positive pair with the query). This was further supplemented with texts by other authors, which created negative pairs with the query text. The task for the model was then to match each query text with the correct target text from the provided set.

5 Results

5.1 Author Verification

Table 3 presents the performance of each model in the tested settings, averaged over three iterations. Only two models — GPT-4o and Claude-3.5 — demonstrated accuracy comparable to the results reported by Huang et al. (2024) for English texts. Both models outperformed the LaBerta-based baseline⁶ in terms of accuracy, with notably high positive predictive values. Although Claude-3.5 did not outperform the baselines in terms of recall and F1 scores, its numbers were higher than those of Mistral and Gemini.

Contrarily to what was expected based on the results yielded by the so-called linguistically-informed prompt reported by Huang et al. (2024),

⁵https://github.com/glsch/sui_generis/blob/main/data/authorship_verification_dataset.pkl.

⁶The model was used without fine-tuning.

Experiment	Parameter	Value
Authorship Verification	Total authors	22
	Pairs per author	30 (15 positive, 15 negative)
	Total pairs	660
	Repetitions	3
	Pairs per iteration	220
	Text length	app. 500 words
Authorship Attribution	Sizes of candidate author sets	5, 10, 15, 22
	Repetitions per configuration	5
	Texts per author	2 (1 query, 1 target)
	Pair types	1 positive, multiple negative
	Text length	app. 500 words and app. 250 words

Table 2: Sampling.

Model	Prompt/Parameters	Accuracy	Precision	Recall	F1
claude-3-5-sonnet-20240620	BASE	72 ($\pm 1\%$)	98 ($\pm 2\%$)	45 ($\pm 4\%$)	62 ($\pm 4\%$)
	BASE_TOPIC_IGNOREANT	68 ($\pm 1\%$)	99 ($\pm 1\%$)	37 ($\pm 4\%$)	54 ($\pm 4\%$)
	HIP	67 ($\pm 4\%$)	98 ($\pm 2\%$)	34 ($\pm 10\%$)	50 ($\pm 11\%$)
	HIP_TOPIC_IGNOREANT	61 ($\pm 3\%$)	100 ($\pm 0\%$)	22 ($\pm 4\%$)	36 ($\pm 5\%$)
	LIMITED	70 ($\pm 5\%$)	99 ($\pm 1\%$)	40 ($\pm 6\%$)	57 ($\pm 6\%$)
	LIMITED_TOPIC_IGNOREANT	67 ($\pm 3\%$)	99 ($\pm 1\%$)	34 ($\pm 1\%$)	51 ($\pm 1\%$)
gemini-1.5-pro	BASE	56 ($\pm 2\%$)	73 ($\pm 6\%$)	18 ($\pm 1\%$)	29 ($\pm 2\%$)
	BASE_TOPIC_IGNOREANT	52 ($\pm 4\%$)	57 ($\pm 11\%$)	21 ($\pm 4\%$)	31 ($\pm 6\%$)
	HIP	57 ($\pm 2\%$)	77 ($\pm 4\%$)	21 ($\pm 5\%$)	33 ($\pm 6\%$)
	HIP_TOPIC_IGNOREANT	54 ($\pm 1\%$)	84 ($\pm 7\%$)	11 ($\pm 2\%$)	19 ($\pm 4\%$)
	LIMITED	56 ($\pm 4\%$)	84 ($\pm 9\%$)	15 ($\pm 4\%$)	25 ($\pm 5\%$)
	LIMITED_TOPIC_IGNOREANT	55 ($\pm 7\%$)	82 ($\pm 11\%$)	12 ($\pm 5\%$)	20 ($\pm 8\%$)
gpt-4o	BASE	78 ($\pm 2\%$)	90 ($\pm 6\%$)	63 ($\pm 1\%$)	74 ($\pm 2\%$)
	BASE_TOPIC_IGNOREANT	70 ($\pm 3\%$)	95 ($\pm 1\%$)	43 ($\pm 3\%$)	59 ($\pm 3\%$)
	HIP	75 ($\pm 1\%$)	88 ($\pm 0\%$)	57 ($\pm 3\%$)	69 ($\pm 3\%$)
	HIP_TOPIC_IGNOREANT	71 ($\pm 1\%$)	96 ($\pm 5\%$)	43 ($\pm 2\%$)	59 ($\pm 3\%$)
	LIMITED	80 ($\pm 2\%$)	89 ($\pm 5\%$)	68 ($\pm 5\%$)	77 ($\pm 3\%$)
	LIMITED_TOPIC_IGNOREANT	70 ($\pm 3\%$)	95 ($\pm 2\%$)	43 ($\pm 1\%$)	59 ($\pm 1\%$)
mistral-large-latest	BASE	56 ($\pm 3\%$)	54 ($\pm 3\%$)	76 ($\pm 5\%$)	63 ($\pm 4\%$)
	BASE_TOPIC_IGNOREANT	54 ($\pm 3\%$)	54 ($\pm 4\%$)	56 ($\pm 3\%$)	55 ($\pm 3\%$)
	HIP	54 ($\pm 6\%$)	53 ($\pm 7\%$)	75 ($\pm 3\%$)	62 ($\pm 6\%$)
	HIP_TOPIC_IGNOREANT	54 ($\pm 5\%$)	53 ($\pm 7\%$)	63 ($\pm 5\%$)	58 ($\pm 6\%$)
	LIMITED	56 ($\pm 1\%$)	54 ($\pm 1\%$)	76 ($\pm 1\%$)	63 ($\pm 1\%$)
	LIMITED_TOPIC_IGNOREANT	53 ($\pm 6\%$)	53 ($\pm 9\%$)	47 ($\pm 7\%$)	50 ($\pm 7\%$)
TF-IDF + Random Forest	char, ngram_range=2,9, max_features=5000	58	59	61	60
LaBerta + Mean pooling + Cosine similarity		69	54	93	68

Table 3: Results for Authorship Verification task on full dataset (22 authors, 5 positive and 5 negative pairs per author in each iteration).

explicit philological and historical features generally deteriorated results compared to the BASE setting for all models except Gemini.

5.2 Author Attribution

Tables 4 and 7 present the results of the authorship attribution task conducted on subsets of 5, 10, and 15 authors, as well as on the full dataset of 22 authors, using text fragments of 500 and 250 words, respectively. Only the GPT-4o model was tested for this task, as it had demonstrated the best performance in the simpler authorship verification setting.

Since multi-class classification is generally more challenging than binary classification, it is unsurprising that GPT-4o did not surpass the LaBerta baseline when the text length was sufficient (500 words). However, in one setting — fragments of 250 words with 5 authors (see Table 7 in Appendix C) — GPT-4o outperformed both baselines. In all other cases, as the number of candidate authors increased, GPT-4o’s performance declined, and at a faster rate than that of the LaBerta baseline.

Consistent with the observations from the authorship verification experiments, explicit instructions regarding philological and histori-

Model	Prompt/Setting	5 Authors		10 Authors		15 Authors		22 Authors	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o	BASE	48	37	32	24	37	28	21	13
	BASE_TOPIC_IGNOREANT	68	62	36	27	37	28	21	12
	HIP	56	48	32	23	35	28	21	13
	HIP_TOPIC_IGNOREANT	44	36	32	22	39	29	17	8
	LIMITED	52	42	28	19	29	22	20	11
	LIMITED_TOPIC_IGNOREANT	56	46	34	25	35	26	14	7
TF-IDF		44	37	26	19	12	7	6	4
LaBerta + Mean pooling + Cosine		72	65	42	34	41	35	36	29

Table 4: Results for Authorship Attribution task on subsets of 5, 10, 15 and 22 (full dataset) authors with fragments of 500 words in terms of Accuracy and Weighted F1. The results of GPT-4o model are compared with several baseline pre-trained models.

cal features had a negative impact on performance. Prompts with fewer constraints, such as `BASE_TOPIC_IGNOREANT` or `LIMITED`, yielded better results. As expected, the length of text fragments had a predictable effect on prediction quality, with accuracy generally decreasing as the texts became shorter (except in the 5-author setting). This suggests that longer fragments provide more information beneficial for authorship attribution.

6 Discussion

The experiment have provided interesting insights into the capabilities of the LLMs and the way how they approach the tasks of authorship verification and attribution.

In Authorship Verification, the strong performance of GPT-4o in the basic setting was largely anticipated due to its advanced capabilities. However, the comparable results achieved by Claude-3.5 are noteworthy, indicating its potential effectiveness in authorship verification tasks.

We were initially concerned about the high performance of the GPT-4o model in the Authorship Verification task, assuming the possibility that parts of our dataset could be simply memorized during the training and merely recalled in our experiment. The decrease in the GPT-4o’s performance observed in the Authorship Attribution task, especially with an increasing number of candidate authors, suggests that the model’s decisions were guided by underlying processes other than reproducing memorized content.

In this respect, the observation that more detailed instructions, crafted by a domain expert based on scholarly tradition, actually deteriorated performance contrasts the performance of the linguistically-informed prompt used by Huang et al. (2024) and is perhaps particularly noteworthy.

While the models are capable of detecting and describing philological features within the texts, this ability does not necessarily translate into accurate predictions. The connection between the features mentioned in HIP to the prediction is much subtler and less straightforward than that of, for example, orthography or punctuation mistakes so successfully used by Huang et al. (2024). This possibly suggests that when a model can leverage its intrinsic knowledge, it achieves better results than when formal instructions are provided for tasks that are resistant to formalization (Ouyang et al., 2022; Liu et al., 2021).

A closer examination of the models’ output⁷ highlights this issue. When explicitly instructed, the models generally perform well in identifying the specified features. For instance, they demonstrated notable “attention” to syntactical patterns such as anaphora (repetition of a word or phrase at the beginning of successive clauses), asyndeton (omission of conjunctions), polysyndeton (repetition of conjunctions), and hyperbaton (disruption of normal word order through the insertion of other words). However, in many cases, the models tend to overinterpret these features, often assuming a deterministic relationship between the presence of such patterns and the final prediction.

For example, when comparing different passages from Leander of Seville, GPT-4o generated the following description of the rhetorical devices in the two texts: “The first text uses rhetorical questions and exclamations to emphasize its points (e.g., ‘O infinita humilitatis documenta!’). The second text, however, relies more on a narrative and descriptive style, with extensive use of quotations from Solomon to build its argument. The rhetori-

⁷All responses are available on GitHub: https://github.com/glsch/sui_generis/blob/main/data/authorship_verification_responses.tsv

cal strategies differ significantly between the two texts.” Although this succinct characterization is adequate, the conclusion reached by the model is incorrect.

Similarly, GPT-4o was perplexed by a discrepancy in two different sermons by Caesarius of Arles, stating: “Text 1 employs a more complex and formal structure, with longer sentences and a higher frequency of subordinate clauses. For example, phrases like *ut modestiae tuae non desit auctoritas, constantiam mansuetudo commendat, iustitiam lenitas temperet* show a sophisticated use of parallelism and balance. Text 2, while still formal, uses shorter sentences and simpler structures. It often employs direct questions and answers, such as *quis est hic, et laudabimus eum?* and *absit, ut desperem hic esse aliquem, immo non aliquem, sed aliquos*. This creates a more conversational tone.” Similar example can be multiplied at random.

6.1 The Role of Semantic Similarity

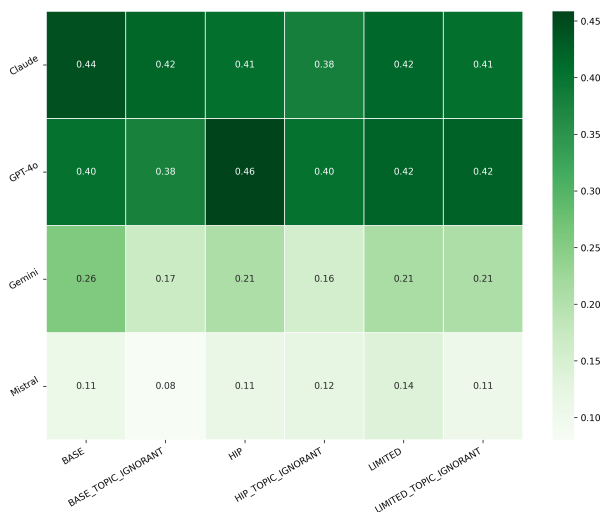


Figure 1: Cosine Similarity Correlation Heatmap by Model and Prompt

Given the high number of such cases, we tried to analyze how semantic similarity influenced the models’s decisions in the authorship verification setting. Figure 1 represents the Pearson correlation coefficient between the cosine similarity of the prompted texts and the correctness of the model’s answers. The responses of the best-performing models, GPT-4o and Claude, seem to align well with semantic similarity across various prompt settings, with only marginal variation. Even when explicitly instructed not to take the content into consideration, the models largely relied on the meaning

of the texts. While writing style and semantics are inherently connected, in authorship analysis, the challenge lies precisely in discerning writing style independently of the subject matter. Our results suggest that LLMs struggle with this distinction, at least in a zero-shot setting.

LLMs are designed to follow human instructions closely, which probably explains why the settings with a lot of explicit guidance show a higher precision. However, the inherently intuitive nature of authorship analysis, especially for short texts, is not easily formalizable, which is in contradiction with what the models are trained for.

When models are given strict prompts, they tend to follow them closely but may overinterpret features, resulting in deteriorated prediction quality. Overprompting seems to limit the models’ ability to leverage their intrinsic knowledge effectively.

It is particularly clear in the case of Gemini. The model seems to have responded positively to provided instructions. With more detailed prompts, the precision of the answers increased, whereas correlation with the semantic similarity diminished. Yet, the instructions — although formulated by a domain expert and synthesized the criteria commonly applied to authorship analysis in the field (Mutzenbecher, 1962; Dolbeau, 2017; Weidmann, 2018) — do not cover all possible stylistic subtleties, limiting its effectiveness. Larger models like GPT-4o and Claude benefit from less constrained prompts, allowing them to apply their extensive intrinsic knowledge more freely and leverage their capability to discern semantic similarities. We suggest that this is the reason, why the LIMITED setting, which gives provides the models with some hint to what to pay attention to and does not constrain them too much, performed that well on both tasks.

7 Conclusion

This study highlights the potential of large language models in performing authorship verification and attribution for Latin texts. The LLMs, particularly GPT-4o, exhibited robust performance, often surpassing traditional baselines. However, our results also highlight the challenges in steering these models’ “decision-making” processes. While LLMs are capable of handling complex linguistic tasks in low-resource historical languages like Latin, there is still significant room for improvement in their interpretability and adaptability to domain-specific nuances. Enhancing their ability

to disentangle style from content without relying overly on thematic similarities is crucial.

By addressing these challenges, we can unlock the full potential of LLMs in philological and historical investigations, contributing valuable tools to the fields of computational linguistics, stylometry, and the digital humanities.

Limitations

This study and the very approach it explores have several limitations one has to keep in mind. First, in zero-shot setting we fully rely in how the models were trained by their creators, and none of the used state-of-the-art LLMs was specifically trained (or tuned) on extensive Latin datasets especially on a rather peculiar and niche task such as authorship analysis. Therefore, experimenting with it might not fully capture the potential of these models. Second, the dataset used in this study is relatively small and, as mentioned in Section 4, is very peculiar from a thematic point of view. While being an interesting benchmark, it might yield observations which are difficult to generalize for texts of other epochs or genres, e.g. Latin poetry, scientific or legal prose. Third, in study, only a very superficial qualitative analysis of the output was performed. Although we present working hypotheses on the models's decision-making based on quantitative observations, the real extent of the relevance of the analysis generated by the models is yet to be determined in a close reading. We intend to investigate this in our future research.

Ethics Statement

This research adheres to the ethical guidelines established by the Association for Computational Linguistics (ACL). We acknowledge the limitations inherent in the use of LLMs, particularly concerning their potential biases and the ethical implications of using proprietary datasets. Care was taken to ensure that the data used did not violate any privacy or copyright concerns. The broader impact of this research is considered in terms of its contribution to the digital humanities, particularly in enhancing the tools available for studying historical texts in low-resource languages. We encourage further research that critically examines the ethical dimensions of applying LLMs to historical and cultural datasets.

Acknowledgements

The work was supported by the ERC Starting Grant Patristic Sermons in the Middle Ages (PASSIM) and ERC Proof of concept grant ManuscriptAI. The PI of both projects is Dr. Shari Boodts, Radboud University (Nijmegen, The Netherlands).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Reduan Achibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2023. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019.
- Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. *Whodunit? Learning to Contrast for Authorship Attribution*. Publisher: arXiv Version Number: 2.
- David Bamman and Patrick J. Burns. 2020. *Latin bert: A contextual language model for classical philology*.
- Pieter Beullens, Wouter Haverals, and Ben Nagy. 2024. *The Elementary Particles: A Computational Stylo-metric Inquiry into the Mediaeval Greek-Latin Aristotle*. *Mediterranea. International Journal on the Transfer of Knowledge*, 9:385–408.
- Janek Bevendorff, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2024. Overview of the “Voight-Kampff” Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings. CEUR-WS.org.
- Shari Boodts and Gleb Schmidt. 2022. Sermon/Homiletics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Patrick Burns. 2023. *How much Latin does ChatGPT “know”?*
- John F. Burrows. 2002. ‘delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

- Jean-Baptiste Camps, Benedetta Salvati, Gonzalo Freijedo, Donghan Bian, Gaëtan Drouet, Eglantine Gaglione, Émilie Guidi, Carolina Macedo, Yaelle Zribi, and Florian Cafiero. 2024. The Authorship of the Works of Chrétien de Troyes: a Stylometric Examination. In *DH Benelux 2024*.
- Thibault Clérice and Anthony Glaise. 2023. [Twenty-One* Pseudo-Chrysostoms and more: authorship verification in the patristic world](#). In *Computational Humanities Research Conference 2023*, Proceedings of the Computational Humanities Research Conference 2022, Paris, France.
- Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. 2023. [Syllabic quantity patterns as rhythmic features for Latin authorship attribution](#). *Journal of the Association for Information Science and Technology*, 74(1):128–141.
- Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, Mirko Tavoni, and others. 2020. L’Epistola a Cangrande al vaglio della computational authorship verification: Risultati preliminari (con una postilla sulla cosiddetta “XIV Epistola di Dante Alighieri”). In *Nuove inchieste sull’epistola a Cangrande: atti della giornata di studi*.
- François Dolbeau. 2017. [Sermons « africains » : critères de localisation et exemple des sermons pour l’Ascension](#). In Gert (Ed) Partoens, A. Dupont, and S. Boodts, editors, *Praedicatio patrum: Studies on preaching in late antique North Africa*, Instrumenta Patristica et Mediaevalia (IPM)_ , pages 9–35. Brepols, Turnhout.
- Anthony Dupont, Johan Leemans, Gert Partoens, and Shari Boodts, editors. 2018. *Preaching in the Patristic Era: sermons, preachers, and audiences in the Latin West*. Number 6 in A new history of the sermon. Brill, Leiden ; Boston.
- Maciej Eder. 2015. [Does size matter? Authorship attribution, small samples, big problem](#). *Digital Scholarship in the Humanities*, 30(2):167–182.
- Maciej Eder. 2017. Short Samples in Authorship Attribution: A New Approach. In *International Conference on Digital Health*.
- Maël Fabien, Esa ú Villatoro-Tello, Petr Motliceck, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing*. ACL.
- Mingcan Guo, Zhongyuan Han, Haoyang Chen, and Haoliang Qi. 2023. A contrastive learning of sample pairs for authorship verification. *Working Notes of CLEF*.
- Rebecca M. M. Hicke and David Mimno. 2023. [T5 meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models](#).
- David I Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28:87–106. Publisher: Springer.
- David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117. Publisher: Oxford University Press.
- Baixiang Huang, Canyu Chen, and Kai Shu. 2024. [Can Large Language Models Identify Authorship?](#) ArXiv:2403.08213 [cs].
- Javier Huertas-Tato, Alvaro Huertas-Garcia, Alejandro Martín, and David Camacho. 2022. [PART: Pre-trained Authorship Representation Transformer](#). ArXiv:2209.15373 [cs].
- Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2024. Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowledge-Based Systems*, page 111867. Publisher: Elsevier.
- Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Andres Karjus. 2023. [Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence](#). Publisher: arXiv Version Number: 1.
- Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *Working notes of CLEF 2020-Conference and Labs of the Evaluation Forum, 22-25 September, Thessaloniki, Greece*, pages 1–14.
- Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, and Walter Daelemans. 2019. Overview of the cross-domain authorship verification task at pan 2019. In *Working Notes Papers of the CLEF 2019 Evaluation Labs*.
- Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.,* pages 1–25.
- Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#).

- Nikos Manousakis and Efstathios Stamatatos. 2023. [Authorship Analysis and the Ending of Seven Against Thebes: Aeschylus' Antigone or Updating Adaptation?](#) *Classical World*, 116(3):247–274.
- Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302):275–309. Publisher: Taylor & Francis.
- Frederick Mosteller and David L. Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer-Verlag New York Inc.
- A. Mutzenbecher, editor. 1962. *Sermonum collectio antiqua, nonnullis sermonibus extravagantibus adiectis (Maximus Taurinensis)*. Number 23 in Corpus Christianorum Series Latina. Brepols, Turnhout. Ed.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
- Mila Oiva, Ksenia Mukhina, Vejune Zemaityte, Tillmann Ohm, Mikhail Tamm, Andres Karjus, Mark Mets, Daniel Chavez Heras, Mar Canet Sola, Helena Hanna Juht, and Maximilian Schich. 2023. [A Framework for the Analysis of Historical Newsreels](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. [Learning Interpretable Style Embeddings via Prompting LLMs](#). ArXiv:2305.12696 [cs].
- Panagiotis Petropoulos. 2023. Contrastive learning for authorship verification using BERT and bi-LSTM in a Siamese architecture. *Working Notes of CLEF*.
- Sahana Ramnath, Kartik Pandey, Elizabeth Boschee, and Xiang Ren. 2024. CAVE: Controllable Authorship Verification Explanations. *arXiv preprint arXiv:2406.16672*.
- Nils Reimers and Iryna Gurevych. 2019. [SentenceBERT: Sentence Embeddings using Siamese BERT-Networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Frederick Riemenschneider and Anette Frank. 2023a. [Exploring Large Language Models for Classical Philology](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.
- Frederick Riemenschneider and Anette Frank. 2023b. [Exploring large language models for classical philology](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*. Association for Computational Linguistics, Toronto, Canada. To appear.
- Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. [Learning universal authorship representations](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Philipp Roelli. 2014. [The corpus corporum, a new open latin text repository and tool](#). *Archivum latinatis medii aevi: Bulletin du Cange*, 72:289–304.
- Mattia Setzu, Silvia Corbara, Anna Monreale, Alejandro Moreo, and Fabrizio Sebastiani. 2024. Explainable authorship identification in cultural heritage applications. *ACM Journal on Computing and Cultural Heritage*.
- Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Tamar Solorio. 2017. [Convolutional neural networks for authorship attribution of short texts](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.
- Efstathios Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *JASIST*, 60:538–556.
- Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21:421.
- Efstathios Stamatatos, Mike Kestemont, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. 2022. Overview of the authorship verification task at PAN 2022. In *CEUR workshop proceedings*, volume 3180, pages 2301–2313. CEUR-WS. org.
- Estathios Stamatatos, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. 2023. Overview of the Authorship Verification Task at PAN 2023. *Working Notes of CLEF*.
- Phillip Benjamin Ströbel. 2022. [Roberta base latin cased v2](#).
- Petter Törnberg. 2024. [Best Practices for Text Annotation with Large Language Models](#). Version Number: 1.

- Martin Volk, Dominic P Fischer, Lukas Fischer, Patricia Scheurer, Phillip Ströbel, Rachele Sprugnoli, and Marco Passarotti. 2024. LLM-based Machine Translation and Summarization for Latin. Publisher: University of Zurich.
- Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. [Can Authorship Representation Learning Capture Stylistic Features?](#) *Transactions of the Association for Computational Linguistics*, 11:1416–1431.
- Clemens Weidmann. 2018. Maximus of Turin. Two Preachers of the Fifth Century. In Anthony Dupont, Shari Boodts, Gert Partoens, and Johan Leemans, editors, *Preaching in the Patristic Era. Sermons, Preachers, and Audiences in the Latin West*, number 6 in A New History of the Sermon, pages 347–372.
- Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. Bert in plutarch’s shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080.

A Intra-model agreement

The intra-model agreement scores reflect the reproducibility and reliability of results across models. High agreement scores, particularly with prompts incorporating topic-ignorance instruction, suggest that these prompts encourage models to make more predictions rather based on features unrelated to the subject matter of the texts.

Table 5 presents the intra-model agreement scores across different prompts for each model comparison. Generally, we observe that models demonstrate higher agreement scores when using prompts with TOPIC_IGNORANCE instruction compared to the generic prompts. The LIMITED_TOPIC_IGNORANT prompt consistently yields higher agreement, especially between Claude and GPT-4o, as well as between GPT-4o and Mistral, suggesting that topic ignorance instructions positively influence intra-model consistency in predictions. Conversely, lower agreement scores are observed between Claude and Mistral, indicating that certain model-prompt pairs may interpret and respond to stylistic cues differently, even when following similar instructions.

B Experiment Settings

Table 6 summarizes the prompts used in the study.

C Authorship Verification 250 words

Table 7 shows the Authorship Attribution results for fragments of 250 words.

Prompt	Claude vs Gemini	Claude vs GPT-4o	Claude vs Mistral	Gemini vs GPT-4o	Gemini vs Mistral	GPT-4o vs Mistral
BASE	79.09	81.06	44.85	68.48	38.03	50.00
HIP	79.09	81.06	44.85	68.48	38.03	50.00
LIMITED	80.91	80.61	43.79	66.67	36.52	51.97
BASE_TOPIC_IGNORANT	72.73	85.61	53.03	69.85	46.82	54.09
HIP_TOPIC_IGNORANT	88.03	84.85	44.55	78.64	42.73	49.09
LIMITED_TOPIC_IGNORANT	81.97	87.12	56.97	78.48	53.48	55.61

Table 5: Intra-model agreement scores across different prompts for model comparisons.

Prompt structure	
System message	
You are an experienced philologist who specializes in post-Classical Latin and has a deep knowledge of Latin patristic literature. Your task is to verify the authorship of texts.	
Task definition	
Authorship Verification	Authorship Attribution
You will be given a pair of texts, and you will have to analyze them in order to decide whether they are written by the same author or not. Importantly, you do not have to guess who the author is, but only decide whether the provided texts are likely to be written by the same person or not.	Given a set of texts with known authors and a query text, determine the author of the query text.
<i>Optional parameter</i>	
TOPIC IGNORANCE As the texts are thematically similar and all of them feature religious, theological, and philosophical content, you should disregard in your decision the topic and content (an additional instruction which can be prepended to any other).	
Guidance levels	
BASE Task definition only, no further guidance provided except for optional TOPIC IGNORANCE .	
LIMITED Base your reasoning on the analysis of the writing style of the input texts.	
HIP (historically-informed prompt) Carry out your analysis by examining the philological and historical elements of the writing style found in the input texts. Consider, but do not limit your analysis to, the following features:	
<ul style="list-style-type: none"> • Morphology: affixes, declination, and verbal endings • Syntax: sentence structure, use of tenses and moods • Rhetorical figures: tropes and figures of speech which alter the ordinary meaning or order of words to produce rhetorical effects or rhythmical patterns. • The use of the Bible: how biblical quotations are introduced, framed, and/or connected to each other • Vocabulary of the text: compound and modal verbs; the words authors use to make evident the structure of the argument as well as various function words (conjunctions, pronouns, interjections, and particles) and the so-called hapax legomena (rare word and expressions) • The tone of the text (moralizing, philosophical, exegetical, high-flown, affectionate, chunky, simplistic, etc.) 	
Human message	
Authorship Verification	Authorship Attribution
<ul style="list-style-type: none"> • Text 1 • Text 2 	<ul style="list-style-type: none"> • Query text • Texts of candidate authors

Table 6: Prompt structure and experiment settings.

Model	Prompt/Setting	5 Authors		10 Authors		15 Authors		22 Authors	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o	BASE	64	55	30	21	31	23	18	11
	BASE_TOPIC_IGNOREANT	56	47	30	24	25	17	18	12
	HIP	60	50	28	19	21	14	22	15
	HIP_TOPIC_IGNOREANT	52	40	28	19	21	15	17	11
	LIMITED	68	61	30	27	24	16	23	17
	LIMITED_TOPIC_IGNOREANT	68	59	32	22	28	19	17	11
TF-IDF		20	15	14	11	13	10	8	5
LaBerta + Mean pooling + Cosine		48	39	66	57	40	31	39	33

Table 7: Results for Authorship Attribution task on subsets of 5, 10, 15 and 22 (full dataset) authors with fragments of 250 words in terms of Accuracy and Weighted F1. The results of GPT-4o model are compared with several baseline pre-trained models.

Enhancing Neural Machine Translation for Ainu-Japanese: A Comprehensive Study on the Impact of Domain and Dialect Integration

Ryo Igarashi
n33t5hin@gmail.com

So Miyagawa
University of Tsukuba
1-1-1 Tennodai
Tsukuba, Ibaraki, Japan
miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

Neural Machine Translation (NMT) has revolutionized language translation, yet significant challenges persist for low-resource languages, particularly those with high dialectal variation and limited standardization. This comprehensive study focuses on the Ainu language, a critically endangered indigenous language of northern Japan, which epitomizes these challenges.

We address the limitations of previous research through two primary strategies: (1) extensive corpus expansion encompassing diverse domains and dialects, and (2) development of innovative methods to incorporate dialect and domain information directly into the translation process. Our approach yielded substantial improvements in translation quality, with BLEU scores 39.06 for Japanese → Ainu and 31.83 for Ainu → Japanese.

Through rigorous experimentation and analysis, we demonstrate the crucial importance of integrating linguistic variation information in NMT systems for languages characterized by high diversity and limited resources. Our findings have broad implications for improving machine translation for other low-resource languages, potentially advancing preservation and revitalization efforts for endangered languages worldwide.

1 Introduction

Ainu is the indigenous language of the Ainu people, who are native to northern Japan, Sakhalin, and the Kuril Islands.

Due to the Japanese government's assimilation policy during the 20th century, the number of people speaking Ainu as their first language drastically declined. Today, UNESCO classifies Ainu as a critically endangered language, and estimates suggest that fewer than ten native speakers remain, all of whom are elderly (Moseley, 2010).

However, there has been a growing focus on revitalizing the Ainu language in recent years. This

development follows the Japanese government's official recognition of the Ainu as an indigenous people, which has led to national funding for Ainu language courses and educational materials (Sato, 2012).

Many Ainu learners speak Japanese as their first language today; thus, practical machine translation is integral to the revitalization of Ainu. However, a previous study by Miyagawa (2023) faced significant challenges, including difficulties in distinguishing between different dialects and challenges in translating everyday conversation.

To address these problems, we carried out the following approaches.

Firstly, we enhanced the corpus. Previous studies' corpora were predominantly biased toward folklore from limited regions. We gathered and digitized resources from various dialects and domains to ensure greater diversity.

We also introduced a novel approach to Ainu-Japanese translation that can distinguish dialects and domains, reducing wording confusion between different regions or contexts.

In this paper, we elaborate on the details of the methodology, present our results, and discuss the implications of our findings that can potentially contribute to the revitalization of Ainu, which may also apply to other low-resource languages.

2 Background and Related Work

In this section, we will review the background of the Ainu language and discuss previous studies.

2.1 The Ainu Language

Ainu is a language isolate with no demonstrable genetic relationship to any other languages, including neighboring languages such as Japanese.

Furthermore, Ainu is a polysynthetic language, where complex words with extensive meanings can be created by combining multiple affixes (Tamura, 2020).

Additionally, Ainu does not have a native writing system, and currently, it is written using the Latin alphabet or Katakana. In particular, the orthography used in the textbook *AKOR ITAK*, published by the former Hokkaido Utari Association, has been broadly accepted by learners and adopted in other publications (Nakagawa, 2006).

2.2 Challenges in Ainu Language Processing

Neural language processing in Ainu faces several significant challenges.

Firstly, the Ainu language is not standardized, leading to regional variations in expressions. These differences are widespread and affect vocabulary, grammar, and pronunciation (Hattori and Chiri, 1960).

Secondly, expressions in Ainu vary significantly depending on context. In Ainu, vocabulary and wording change based on whether the language is used in storytelling, such as folklore narration, or in everyday conversation. One notable difference is the use of personal affixes reflecting logophoricity (Bugaeva, 2008). For example, in conversation, actions of a speaker are marked by a first-person prefix *ku=*, but in folklore, it often changes to a fourth-person affix *a=* or *=an*. Conversely, using *a=* or *=an* in conversation indicates a quotation or an inclusive "we" (Nakagawa, 2011).

Lastly, the availability of corpora for Ainu language processing is extremely limited. Although institutions including the National Institute for Japanese Language and Linguistics (NINJAL) and the National Ainu Museum have made efforts to collect and digitize some corpora, the overall quantity remains insufficient. Moreover, most of these corpora focus on folklore from specific regions, which does not adequately capture the full diversity of the Ainu language.

Therefore, it is essential to expand the corpus to include resources from various domains and regions and to enhance the machine translation model's ability to handle ambiguities. These steps are crucial for improving the performance of natural language processing in Ainu.

2.3 Previous Work in Ainu-Japanese Machine Translation

Ptaszynski et al. (2013) proposed an initial implementation of rule-based Ainu-Japanese machine translation. This system internally uses a part-of-speech tagger based on a Hidden Markov Model,

replacing each Ainu word with its Japanese equivalent. This approach works well for Japanese, where word order is closely aligned with Ainu.

Furthermore, Miyagawa (2023) experimented with Transformer-based neural machine translation in Ainu-Japanese, achieving a BLEU score of 32.90 for Japanese-to-Ainu and 10.45 for Ainu-to-Japanese, the highest reported to date for this task.

However, it is important to note that both studies trained their models using folklore from a specific region, which limits their applicability to conversations or other dialects.

3 Methodology

This section details our methodology for selecting materials in the corpus, preparing them in a format suitable for machine learning, and conducting the training process.

3.1 Corpus

As mentioned above, most digitized corpora currently available are lacking in diversity, primarily containing folklore from limited regions. To address this, we collected additional resources to enhance the diversity of our corpus.

In addition to already digitized resources, we established the following criteria to prioritize resources for digitization:

- **Writing System:** We exclusively selected resources written in the Latin alphabet according to the *AKOR ITAK* orthography. Although Katakana is also widely used, we chose not to include it due to its lower accuracy in Optical Character Recognition (OCR), which is essential for converting printed text into a digital format. Additionally, while it is possible to convert Latin scripts into Katakana, the reverse is not feasible, as Katakana does not distinguish between phonemes such as *i* and *y* or *u* and *w*. Moreover, Katakana does not clearly indicate the boundaries between personal affixes and other word components.
- **Wider Variety:** We prioritized resources that contain modern texts, conversations, and less documented dialects to ensure a comprehensive and representative dataset that reflects the full range of linguistic diversity in Ainu.

As a result of an extensive review of available books, websites, textbooks, dictionaries, and peri-

odicals, we successfully collected a diverse set of resources, which are detailed in Appendix A.

3.2 Preprocessing

Most of the resources we selected are not machine-readable. Even when they were available as PDF files, we needed to establish the correspondence between Ainu sentences and their counterpart Japanese translations. We addressed this issue through the following steps.

Firstly, we scanned the printed materials and used the Cloud Vision API¹ to perform OCR. This allowed us to obtain machine-readable text data along with metadata, including dimensions and coordinates, similar to the metadata structure typically found in PDFs. Although the Cloud Vision API does not officially support the Ainu language, it has basic recognition capabilities for the Latin alphabet. Therefore, we adapted it with additional validation, as described in the following step.

Next, we developed a Node.js script to establish the correspondence between Ainu text and Japanese translations, aligning the parallel corpus. This script utilized the dimensions and coordinates obtained in the previous step. We adjusted the threshold configurations for alignment based on the layouts of each resource.

Finally, we validated the obtained parallel corpus. As expected, the Cloud Vision API, which does not support the Ainu language, produced various recognition errors. These errors included, but were not limited to, incorrect recognition of personal affixes (e.g., *k=arpa* recognized as *karpa*) and issues with single-character words being joined with their sibling words (e.g., *ye p* recognized as *yep*). To address these errors, we validated the recognized text using the customizable open-source spell checker Code Spell Checker², with a word list extracted from the dictionaries available on the Ainu Language Archive³. We then conducted a thorough manual review to correct any remaining recognition errors.

3.3 Format

In this section, we outline the corpus format developed for this study. While alternative methods could offer advantages from the perspective of indexing, we prioritized simplicity, given that our

primary objective was developing a neural machine translation model.

3.3.1 Domains

One of the significant challenges we faced was categorizing the domains of the collected resources.

We initially considered Nowakowski et al.'s (2018) approach, which classified existing corpora into 15 distinct genre types. However, some of our newly collected resources lacked the necessary detail for categorization within the context of Ainu folklore. Accurate classification would have required a deep understanding of Ainu language and culture.

Therefore, we opted against a nuanced categorization of all resources and instead adopted a more straightforward and objective approach: utilizing the personal affix predominantly used in a resource as a proxy for domain categorization. This approach allowed us to automatically classify the resources into two mutually exclusive classes: first-person and fourth-person.

3.3.2 Dialects

Designing maintainable classes for dialects also presented significant challenges due to the lack of consensus on how to classify the various Ainu dialects.

Hattori and Chiri (1960), a pioneering study of Ainu dialects, lists 19 dialects, while a subsequent survey by Asai (1974) expanded this to 21, adding three additional dialects.

As a result, the dialect names used in different resources are inconsistent. For instance, the Ainu Times does not distinguish between the Saru and Chitose regions, instead referring to them collectively as "Saru-Chitose".

Given these inconsistencies, we ultimately decided to retain the dialect names as they were listed in each resource. Consequently, our corpus includes instances labeled as "Saru", "Saru-Chitose", and "Western Hokkaido". While this approach may result in overlapping categories, we did not consider this issue critical, as we hypothesized that modern language models would be capable of discerning the similarities and differences between these dialects.

Finally, we consolidated all the data into the format shown in Table 1 and compiled it using the Hugging Face Datasets library. This process resulted in the creation of a novel corpus (Table 2)

¹<https://cloud.google.com/vision>

²<https://cspell.org/>

³<https://ainugo.nam.go.jp/>

comprising approximately 1.2 million words and 4.2 million characters.

3.4 Normalization

We applied minimal normalization, which involved removing diacritics (e.g., *húre* → *hure*), linking symbols (e.g., *or_un* → *or un*), and footnote markers inserted by editors.

3.5 Training

In this section, we provide a detailed explanation of the training setup.

3.5.1 Model

Lee et al.’s (2022) prior research demonstrated the effectiveness of multilingual sequence-to-sequence language models, such as mT5 (Xue et al., 2020) and mBART (Chipman et al., 2022), for handling machine translation tasks in low-resource languages. For the present study, we selected mT5 due to the extensive research supporting its performance.

Firstly, to reduce costs, we trained the *mt5-small* model, which features a relatively small number of parameters, using various task prefixes, as described in the following section. This preliminary step allowed us to confirm that including metadata, such as dialects and domains, contributes to improved performance.

Next, using the most effective task prefix identified in the previous step, we trained and compared the *mt5-base* and *t5-base* models. This comparison aimed to assess both the effectiveness of the multilingual model and the impact of the number of parameters on performance.

Finally, we conducted additional training with the *mt5-base* model using only specific domain-dialect pairs to evaluate the effectiveness of mixing different types of resources within the corpus. In this step, we focused on folklore and conversations from the Saru, Chitose, Shizunai, and Horobetsu regions, where relatively abundant data are available.

For practical reasons, we trained both the Ainu-to-Japanese and Japanese-to-Ainu translation tasks within the same model.

3.5.2 Task Prefix

T5 and mT5 support multi-task learning by embedding tokens known as task prefixes in an input sequence. This approach enhances overall perfor-

mance by allowing the model to differentiate between various task types (Raffel et al., 2020).

We hypothesized that applying this method to Ainu could improve machine translation performance by effectively disambiguating subtle linguistic differences across dialects and domains.

To test this hypothesis, we conducted experiments using the four variations of language names shown in Table 3.

Here, dialect refers to the specific dialect of the resource, and domain refers to the predominant personal affix used in the text. For example, if translating an Ainu folklore text written in the Saru dialect to Japanese, the P_{both} prefix would be translate Ainu (Saru, fourth) to Japanese. For translations in the opposite direction, the task prefix would be adjusted accordingly to translate Japanese to Ainu (Saru, fourth).

3.5.3 Settings

The training process was conducted using the Hugging Face Transformers library. To orchestrate the infrastructure and ensure consistent metric measurement within the same environment, we developed a training pipeline using Google Cloud’s Vertex Pipelines.

Hyperparameter tuning was performed using Vertex AI’s Hyperparameter Tuning Job. We utilized 10% of the dataset to perform a grid search for optimal hyperparameters, as outlined in Table 4.

3.5.4 Evaluation

We employed stratified sampling (Japkowicz and Stephen, 2002) by dialect-domain pairs to split the dataset proportionally. Specifically, 10% of the corpus was allocated as the evaluation set, while the remaining 90% was used for training.

Evaluation metrics were calculated using the BLEU score (Papineni et al., 2002), with SacreBLEU, a commonly used library in prior research, employed for its computation. For Japanese-to-Ainu translations, we used the default 13a tokenizer, and for Ainu-to-Japanese translations, we used the *ja-mecab* tokenizer.

4 Results

In this section, we elaborate on the results of the experiments.

4.1 Task Prefixes

Here, we examine how performance varies depending on different task prefixes. Table 5 shows the

Name	Type	Description
book	str	Book title
title	str	Title of the text
domain	enum of "first" and "fourth"	Type of personal affix of the speaker
author	Option[str]	Author of the text (if known)
dialect	Option[str]	Dialect of the text (if known)
text	str	Sentence in Ainu
translation	str	Translation in Japanese

Table 1: Corpus format

Title	Type	Words	Characters
The Ainu Language Archive	Web	600,770	2,107,984
The Ainu Times	Book	148,843	519,040
Collection of Ainu Oral Literature	PDF	135,649	492,484
ILCAA Online Text of Ainu Collected by Suzuko Tamura	Web	95,379	299,630
A Glossed Audio Corpus of Ainu Folklore	Web	76,550	243,696
Dictionary of the Mukawa dialect of Ainu	Web	66,386	247,637
Ainu textbooks by The Foundation for Ainu Culture	PDF	25,067	84,905
Bulletin of the Hokkaido Ainu Culture Research Center	PDF	14,724	48,092
A Topical Dictionary of Conversational Ainu	Web	13,831	49,776
Ainu Shin’yōshū	Book	10,364	38,153
New Express Plus Ainu-go	Book	4,418	14,812
Learning Ainu Language by Listening to Kamuy Yukar	Book	3,028	11,177
AKOR ITAK	Book	2,005	5,903
Total		1,197,014	4,163,289

Table 2: The Ainu language resources. We treated personal affixes as separate words and excluded line breaks and whitespace from the character count.

Label	Language Name in Task Prefix
P_{none}	Ainu
P_{dialect}	Ainu (dialect)
P_{domain}	Ainu (domain)
P_{both}	Ainu (dialect, domain)

Table 3: The list of strings we used as a language name for the Ainu language.

BLEU scores for each task prefix.

With P_{none} , we observed a BLEU score of 29.89 for Ainu-to-Japanese and 32.24 for Japanese-to-Ainu, which does not significantly differ from the results reported by Miyagawa (2023).

However, with P_{domain} , we observed a subtle performance improvement, with scores of 29.93 for Ainu-to-Japanese and 32.70 for Japanese-to-Ainu.

With P_{dialect} , performance improved significantly for Japanese-to-Ainu, with a BLEU score of 35.94. We also observed a slight improvement for Ainu-to-Japanese, with a score of 30.40.

Finally, with P_{both} , we achieved the highest performance, with BLEU scores of 30.70 for Ainu-to-Japanese and 36.25 for Japanese-to-Ainu.

4.2 Models

We also experimented with different sequence-to-sequence models to determine which one performs best. Given that P_{both} was proven to be the most effective task prefix, all models were trained using this prefix. Table 6 shows the BLEU scores for each model.

With mt5-base, performance improved for both translation directions, achieving BLEU scores of 31.83 for Ainu-to-Japanese and 39.06 for Japanese-to-Ainu, making it the best-performing model among all the models tested.

In contrast, the t5-base model failed to produce practical translation results, with BLEU scores of 0.00 for Ainu-to-Japanese and 0.01 for Japanese-to-Ainu. As these scores indicate, the model generated nothing but nonsensical text.

Parameter	Value
Framework	Hugging Face Transformers (v4.40.1)
Infrastructure	Google Cloud Vertex AI
Hardware	a2-highgpu-1g instance with NVIDIA A100 GPU
Scheduler	Linear scheduler with 6% warm-up steps
Learning rate	Maximum of 5.0×10^{-5}
Optimizer	AdamW with weight decay of 1.0×10^{-3}
Context size	128 tokens
Batch size	16 with gradient accumulation every 2 steps
Training duration	Maximum of 20 epochs with early stopping (patience=3)

Table 4: Training settings

Task Prefix	ain→ja	ja→ain
P_{none}	29.89	32.24
P_{domain}	29.93	32.70
P_{dialect}	30.40	35.94
P_{both}	30.70	36.25

Table 5: Ainu-Japanese translation performance for each task prefix

Model	ain→ja	ja→ain
t5-base	0.00	0.01
mt5-small	30.70	36.25
mt5-base	31.83	39.06

Table 6: Ainu-Japanese translation performance for each model

4.3 Performance for Each Domain and Dialect

We also evaluated the performance of the mt5-base model across different domains and dialects. Table 7 compares the performance of the mixed corpus model with that of a model trained exclusively on specific domain-dialect pairs.

Across all classes, the model trained on the mixed corpus consistently outperformed the model trained on individual classes. Notably, we observed significant improvements in classes with smaller datasets, such as conversations in the Chitose or Horobetsu dialects.

5 Discussion

This section discusses how multilingual pre-trained models, domain, and dialectal variations impact Ainu machine translation.

5.1 Effectiveness of Domain and Dialect

This study confirmed that incorporating dialectal information significantly improves translation per-

formance. This improvement is likely due to the linguistic variations that exist across different regions. Notably, greater performance gains were observed in the Japanese-to-Ainu translation. This difference may be due to the ambiguity in determining which dialect to use as the target when translating from Japanese. By specifying the target dialect, this ambiguity is resolved, allowing the model to produce more accurate translations and resulting in improved performance.

Here is an example of translating "I go to the mountain to pick mushrooms" (*Watashi wa kinoko o tori ni yama e ikimasu.*) to different dialects. Note that the model correctly used the appropriate wording for each dialect:

Saru: karus ku=kar kusu **ekimne** k=**arpa**.

Tokachi: karus ku=kar kusu **ekimun** ku=**oman**.

Our study also found that the inclusion of domain metadata led to performance improvements, although these gains were less pronounced compared to those achieved with dialectal information. One possible reason for this difference could be the complexity of the vocabulary and unique expressions found in folklore, which may have posed challenges for the model.

Here is an example of translating "I want to eat with my friend" (*Tomodachi to issho ni shokuji shitai.*) in different domains. Note that the model correctly adjusted the personal affixes according to the domain:

Folklore: a=utari turano ipe=**an** rusuy.

Conversation: k=utari turano **ku**=ipe rusuy.

We believe this approach could also be applicable to other languages that lack standardization, especially endangered languages with context and dialect variations. For instance, the Ryukyuan languages, characterized by an extensive politeness

Dialect	Domain	Words	Exclusive Corpus		Mixed Corpus	
			ain→ja	ja→ain	ain→ja	ja→ain
Saru	Conversation	25,506	23.05	33.03	35.47	42.94
	Folklore	527,728	24.48	32.14	28.86	33.31
Shizunai	Conversation	24,403	22.37	37.57	37.48	50.74
	Folklore	233,134	36.35	44.16	38.81	47.31
Chitose	Conversation	6,487	13.51	13.92	70.97	70.96
	Folklore	15,664	12.02	27.51	36.88	42.36
Horobetsu	Conversation	4,382	0.22	0.36	89.19	80.89
	Folklore	10,364	1.23	2.19	30.56	38.21

Table 7: Ainu-Japanese translation performance metrics for each domain and dialect

system and numerous dialects influenced by the archipelagic geography, present a similar challenge. Embedding politeness levels and regional information in the task prefix could improve MT performance for these languages by providing more accurate and contextually appropriate translations. The model might better manage linguistic nuances and variability by explicitly incorporating such meta-data, enhancing translation accuracy.

5.2 Advantages of Multilingual Pre-Trained Models

Building on previous studies, our research confirms the applicability of multilingual pre-trained language models for Ainu-Japanese translation. This finding supports the use of these models for low-resource languages and demonstrates the potential of transfer learning, even for a language isolate such as Ainu.

However, these models should not be considered a universal solution, as they have several drawbacks. One significant disadvantage is the large number of parameters required by mT5, which has a considerably larger vocabulary size to handle tokens from multiple languages. While this enables them to provide strong baseline performance across various language tasks, it also results in excessive parameters for specific tasks, such as translation between particular languages. This leads to inefficiencies and requires substantial computational resources for both training and inference.

Additionally, the Sentencepiece tokenizer (Kudo and Richardson, 2018) used in mT5 was not specifically trained on Ainu texts, leading to suboptimal tokenization. For example, basic sentences such as *irankarapte tanto sirpirka wa* are tokenized into eleven separate tokens, with even fundamental words including *pirka* being split unnecessarily:

['_ir', 'ankara', 'pte', '.', '_', 'tanto', '_sir', 'pirk', 'a', '_wa', '.']

5.3 Impact of Mixing Multiple Dialect-Domain Classes on Model Performance

Our study found that training a model by mixing multiple domain-dialect classes and using task prefixes to distinguish them results in higher performance compared to training on a single domain-dialect class. This effect is particularly pronounced in classes with limited resources. This finding suggests that there is shared grammar or vocabulary among different classes, which a language model can leverage to enhance performance when needed.

This finding could also be valuable for other endangered languages where collecting more materials for a specific dialect is impractical. Our results demonstrate that MT performance can still be enhanced by incorporating resources from other dialects and distinguishing them using task prefixes. This approach allows the model to make use of shared linguistic features across dialects, effectively broadening the usable data pool and compensating for individual dialect resource limitations.

6 Conclusion

This study demonstrates the effectiveness of incorporating dialect and domain information in NMT systems for low-resource languages with high linguistic variation. By expanding the corpus and leveraging task prefixes to provide contextual information, we achieved significant improvements in Ainu-Japanese translation performance.

Our research contributes to the broader field of low-resource language NMT by:

1. Highlighting the importance of diverse, well-annotated corpora.

2. Demonstrating the potential of integrating linguistic metadata into the translation process.
3. Providing insights into the challenges and strategies for addressing languages that lack standardization and exhibit high variation.

As we continue to refine and expand these methods, we move closer to achieving effective machine translation for all languages, irrespective of their resource status. This work not only advances NMT research but also contributes to broader language preservation and revitalization efforts, providing new tools and methodologies for engaging with endangered languages.

Limitations

While our research achieved the highest scores to date in Ainu-Japanese machine translation, several limitations affect the generalizability and performance of our model.

Limited Corpus

The most significant limitation is the insufficient amount of the Ainu language data. Despite our efforts to digitize and format the most comprehensive Ainu corpus currently available, the dataset remains too limited for extensive machine translation training.

This scarcity of data is particularly noticeable in the lack of folklore from regions other than Saru and Shizunai, as well as a shortage of conversational resources across all dialects.

Additionally, some existing resources could not be utilized due to inconsistencies in writing systems. Developing a model that can convert between different writing systems may help address this issue. While data augmentation methods, such as back-translation, could be used to expand the corpus, their effectiveness is likely limited, as most existing Ainu resources come with Japanese translation.

Furthermore, we decided not to make our collected corpus publicly available due to copyright restrictions. This limitation poses challenges for performance comparisons in future research. A model trained exclusively on specific domain-dialect pairs could potentially achieve a higher BLEU score than our model, but this would not necessarily indicate superior performance across the broader spectrum of the Ainu language. Establishing more

consensus on the digitization and use of copyright-protected works, particularly for endangered languages, could help address this issue and facilitate broader research efforts.

Fine-Grained Dialects and Domains

In this study, we employed a simplified approach to classify domains and dialects. Consequently, our model cannot perform translations that target more specific regions or dialects. For example, there are different types of folklore, such as *yukar* and *uwepeker*, both of which are narrated using the same fourth-person affix. Our current approach does not differentiate between these types.

Future research would benefit from a more finely annotated corpus, particularly with respect to capturing subtle differences in dialects or domains.

References

- Toru Asai. 1974. *Classification of dialects: Cluster analysis of Ainu dialects*. Institute for the Study of North Eurasian Cultures, Hokkaido University.
- Anna Bugaeva. 2008. Reported discourse and logophoricity in southern hokkaido dialects of ainu. *GENGO KENKYU (Journal of the Linguistic Society of Japan)*, 133:31–75.
- Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. 2022. mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.
- Shirō Hattori and Mashiho Chiri. 1960. Ainu-go shuhōgen no kiso goi tōkeigakuteki kenkyū [a lexicostatistical study of basic vocabulary in Ainu dialects]. *Minzokugaku kenkyū*, 24(4):307–342.
- Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.
- Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.
- En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? *arXiv preprint arXiv:2203.08850*.
- So Miyagawa. 2023. Machine translation for highly low-resource language: A case study of Ainu, a critically endangered indigenous language in northern Japan. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing*

for *Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 120–124.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Hiroshi Nakagawa. 2006. Ainu-jin ni yoru ainu-go hyōki e no torikumi [Ainu people's efforts to write the Ainu language]. *Writing unwritten languages*, pages 1–44.

Hiroshi Nakagawa. 2011. Ainu no shin'yō ni okeru jojutsu-sha no ninshō [The person of the narrator in Ainu mythology]. *Northern Language Studies*, 1:139–156.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2018. A proposal for a unified corpus of the Ainu language. *IPSJ SIG Tech. Rep.*, 237:1–6.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Michal Ptaszynski, K Mukaichi, and Yoshio Momouchi. 2013. Nlp for endangered languages: Morphology analysis, translation support and shallow parsing of Ainu language. In *Proceedings of the 19th Annual Meeting of The Association for Natural Language Processing, Nagoya, Japan*, pages 12–15.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Tomomi Sato. 2012. Ainu-go no genjō to fukkō [The state of Ainu language and revitalization]. *Gengo Kenkyū*, 142:29–44.

Suzuko Tamura. 2020. *Ainu-go no sekai [The World of Ainu Language]*, shinsō fukyū-ban edition. Yoshikawa kōbunkan.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

A Full List of Resources

In this section, we list the resources included in our corpus.

1. **A Glossed Audio Corpus of Ainu Folklore⁴**: A resource provided by NINJAL, comprising 30 stories of Ainu oral literature from

⁴<https://ainu.ninjal.ac.jp/folklore/>

the Chitose and Saru regions. Each entry includes part-of-speech classification and English translations.

2. **A Topical Dictionary of Conversational Ainu⁵**: A practical phrasebook of the Ainu language, originally compiled by Shozaburo Kanazawa in 1898. For this study, we utilized a transcribed version hosted by NINJAL.
3. **Ainu Shin'yōshū (revised by Tatsumine Katayama)**: A collection featuring modern Japanese translations of Yukie Chiri's *Ainu Shin'yōshū*, alongside texts transcribed in the modern writing system. It comprises 13 yukar tales. Although Hideo Kirikae also produced a modern revision, we adopted Katayama's edition because it includes symbols indicating the boundaries of personal affixes, which follows *AKOR ITAK* orthography.
4. **Ainu Textbooks by The Foundation for Ainu Culture⁶**: A series of textbooks for learning the Ainu language, published by The Foundation for Ainu Culture. Available as PDFs on their website, these materials cover eight dialects across three proficiency levels.
5. **AKOR ITAK**: A textbook published by the Hokkaido Utari Association in 1994. It features the Ainu language in various dialects and explains basic vocabulary, rituals, and folklore. It is also known for proposing an Ainu orthography for the Latin alphabet and Katakana. For this study, we used only the grammar lesson sections containing conversation examples.
6. **Bulletin of the Hokkaido Ainu Culture Research Center⁷**: A research bulletin featuring papers with transcriptions in both Ainu and Japanese. For this study, we excerpted articles from issues 9, 10, 11, 12, and 17.
7. **Collection of Ainu Oral Literature⁸**: A collection of oral literature from Biratori Town, compiled by Shigeru Kayano and transcribed

⁵<https://ainu.ninjal.ac.jp/topic/dictionary/en/>

⁶https://www.ff-ainu.or.jp/web/potal_site/details/post.html

⁷https://ainu-center.hm.pref.hokkaido.lg.jp/05_001.htm

⁸<https://nibutani-ainu-museum.com/culture/language/story/>

by researchers at Chiba University as part of a research project by the Agency for Cultural Affairs. This collection is available on the website of the Nibutani Ainu Culture Museum.

8. **Dictionary of the Mukawa dialect of Ainu⁹**: A phrasebook of the Mukawa dialect of Ainu, compiled by Tatsumine Katayama and available in CSV format on the website of the Graduate School of Humanities and Social Sciences at Chiba University.
9. **ILCAA Online Text of Ainu Collected by Suzuko Tamura¹⁰**: A website featuring online texts of the Ainu language with Japanese translations, collected by Suzuko Tamura. Audio recordings accompany each sentence.
10. **Learning Ainu Language by Listening to Kamuy Yukar**: A textbook by Hiroshi Nakagawa, published by Hakusuisha, focusing on Ainu grammar through Kamuy Yukar from the Chitose region.
11. **New Express Plus Ainu-go**: Another textbook by Hiroshi Nakagawa, published by Hakusuisha, focusing on everyday conversations from the Saru region and explaining Ainu grammar.
12. **The Ainu Language Archive¹¹**: A website maintained by the National Ainu Museum, providing the largest corpus of Ainu language texts alongside their Japanese translations.
13. **The Ainu Times**: A periodical published by the Ainu Language Pen Club, consisting of essays, news articles, and various other writings. Since its inception in 1997, it has released 80 issues. Articles are contributed by volunteers, with each piece indicating the specific dialect used at the conclusion of the text. In this study, we utilized 71 issues from No. 3 to No. 80.

⁹<http://itelmen.placo.net/Ainu-archives/mukawa/>

¹⁰<https://online-resources.aa-ken.jp/resources/detail/IOR000018>

¹¹<https://ainugo.nam.go.jp/>

Exploring Large Language Models for Qualitative Data Analysis

Tim Fischer, Chris Biemann

Language Technology Group, Department of Informatics, University of Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

This paper explores the potential of Large Language Models (LLMs) to enhance qualitative data analysis (QDA) workflows within the open-source QDA platform developed at our university. We identify several opportunities within a typical QDA workflow where AI assistance can boost researcher productivity and translate these opportunities into corresponding NLP tasks: document classification, information extraction, span classification, and text generation. A benchmark tailored to these QDA activities is constructed, utilizing English and German datasets that align with relevant use cases. Focusing on efficiency and accessibility, we evaluate the performance of three prominent open-source LLMs - Llama 3.1, Gemma 2, and Mistral NeMo - on this benchmark. Our findings reveal the promise of LLM integration for streamlining QDA workflows, particularly for English-language projects. Consequently, we have implemented the *LLM Assistant* as an opt-in feature within our platform and report the implementation details. With this, we hope to further democratize access to AI capabilities for qualitative data analysis.

1 Introduction

The Discourse Analysis Tool Suite (Schneider et al., 2023) is a platform developed at our university to empower Digital Humanities (DH) researchers in conducting qualitative data analysis (QDA). Developed collaboratively and tailored to the specific needs of DH scholars, the platform democratizes access to machine learning methods, enabling non-experts to manage and analyze large-scale, unstructured, multi-modal data effectively.

While the platform's overarching design is rooted in Grounded Theory-based research (Strauss and Corbin 1990, Strauss et al. 1996), its versatile features support various disciplines. Within the core QDA workflow on our platform, researchers engage in a dynamic process of structuring their

data and conducting in-depth qualitative analysis. This involves organizing and categorizing documents through metadata assignment and creating a tag taxonomy, which is utilized for classifying documents. Simultaneously, they dive into the nuances of the material, developing hierarchical code taxonomies to annotate relevant text passages and capturing their insights through memos.

While tasks like metadata extraction and document classification can be repetitive and time-consuming, qualitative analysis tasks such as annotation, paraphrasing, and summarization are demanding. The potential for AI assistance to streamline and enhance these diverse workflows motivates our exploration of Large Language Models (LLMs). Hence, this work aims to assess how effectively LLMs can support users in QDA tasks and ultimately boost their efficiency and productivity. To this end, we identify four core NLP tasks embedded in our QDA platform's core functionalities: 1) document classification, 2) document information extraction, 3) span classification, and 4) text generation. We then curate datasets that closely align with real-world use cases regarding domain and tasks, focusing on English and German. Subsequently, we evaluate the performance of three state-of-the-art open-source LLMs, Llama 3.1, Gemma 2, and Mistral NeMo, on this benchmark.

Our findings show the promising potential of LLM integration within the Discourse Analysis Tool Suite (DATS), particularly for English projects. Consequently, we implement the *LLM Assistant* as an opt-in feature for English projects, paving the way for further enhancements and expansions. Contributions of this paper are:

1. We articulate the envisioned AI-assisted workflow within our platform, highlighting user needs and requirements.
2. We design a benchmark tailored specifically to common QDA tasks within our platform.
3. We evaluate open-source LLMs on it.

4. We report on their integration into our tool.

This work represents our first step towards automating and providing assistance for various common tasks in our QDA platform using LLMs. We aim to facilitate more efficient, insightful qualitative data analysis by augmenting researchers' capabilities with LLM assistance.

2 Related work

QDA Platforms and AI Integration Several prominent platforms have emerged in the realm of qualitative data analysis software, each offering distinct functionalities to researchers. Some platforms have taken notable steps towards incorporating AI-powered features into their workflows.

CATMA (Gius et al., 2022) is a versatile QDA tool focusing on text and image analysis. It currently lacks built-in AI capabilities.

Known for its comprehensive approach to qualitative and mixed-methods research, MAXQDA¹ has introduced "MAXQDA AI Assist," offering AI-driven features like summarization, paraphrasing, and concept explanation.

A robust platform for qualitative data analysis, NVivo's² latest beta version is actively integrating AI functionalities, including thematic coding, sentiment analysis, and text summarization.

Atlas.ti³ is recognized for its visual and network-based analysis tools. The platform incorporates AI with existing features like code suggestions, sentiment analysis, summarization, and entity recognition powered by OpenAI's GPT models.

Notably, AI-powered features within these QDA platforms are currently only found in paid versions. They are realized by sending data to third-party providers, potentially leading to data protection issues. In contrast, our open-source QDA platform aims to democratize access to state-of-the-art AI capabilities, making advanced functionalities freely available to researchers across disciplines. It can be run in-house if required.

LLM Benchmarks While several prominent general LLM benchmarks like MMLU (Hendrycks et al., 2021), SuperGLUE (Wang et al., 2019), BIG-bench (Srivastava et al., 2023), HELM (Liang et al., 2023), and MTEB (Muennighoff et al., 2023) exist, they often lack a direct connection to specific real-world applications, including qualitative data analysis. They may cover a broad range of tasks

¹ <https://maxqda.com>

² <https://nvivo.de/>

³ <https://atlasti.com>

but not necessarily those most relevant to QDA workflows. In contrast, Ziems et al. (2024) explores the potential of LLMs to transform Computational Social Science (CSS) by evaluating their zero-shot performance on a range of English CSS tasks. Their extensive evaluation, focused on taxonomic labeling and free-form coding, highlights LLMs' potential to augment CSS research as zero-shot data annotators, strongly motivating our work. Still, many benchmarks prioritize English data.

Our benchmark is constructed to be relevant to the tasks and data encountered in QDA, incorporating datasets that approximate real-world scenarios within our platform. Focusing on document classification, information extraction, span classification, and text generation in English and German, we aim to gather insights that can directly inform the effective integration and utilization of LLMs in qualitative research.

LLMs and QDA Rasheed et al. (2024) explores the potential of LLMs to serve as data analysts in qualitative research within Software Engineering. Their approach employs a multi-agent model where each LLM agent performs specific research-related tasks, such as interpreting textual data and interview transcripts, to automate common qualitative analysis processes. Their findings suggest that LLMs can significantly accelerate data analysis, allowing researchers to handle larger datasets efficiently, which further motivates this work.

3 Envisioned workflow

This section presents an illustrative excerpt of a qualitative data analysis workflow inspired by project partners who actively utilize the Discourse Analysis Tool Suite. This scenario highlights potential areas where AI-powered automation enhances productivity.

Imagine Alice, a researcher aiming to analyze local companies across various industries, focusing on their societal impact and challenges. She initiates semi-structured interviews with CEOs, stating the current date, introducing herself, and then inquiring about company details (e.g., size, sales volume) and the interviewee's background (e.g., name, age, position) before asking her research questions. She records these interviews with her smartphone.

After collecting data, Alice starts the qualitative analysis process within our QDA platform. She creates a new project, defines document tags for industry categorization, and establishes metadata

fields like "Interview Date," "Company Size," and "Partner Name" to capture crucial interview details. Upon uploading her recorded interviews, our platform currently utilizes Whisper (Radford et al., 2022) to generate automatic transcripts.

After the automatic pre-processing, Alice is presented with interview transcripts, which are now automatically tagged by industry and partially populated with metadata. The AI-powered system suggested tags and metadata values based on the interview content, which Alice verifies and completes with the help of an intuitive UI.

Having an organized document collection, Alice starts the qualitative annotation. She constructs a code system aligned with her research questions that incorporates codes like "Social Impact" and fine-grained sub-codes for "Problems". Next, she activates the auto-coding feature, and the AI-powered system suggests relevant text annotations. An interface allows her to review these suggestions.

While reviewing, Alice notices occasional disfluencies like repetitions and filler words, typical of verbatim transcriptions. She selects a disfluent passage, activates the paraphrasing feature, and is presented with an AI-generated suggestion. After minor edits, she approves the improved version. Similarly, she employs the automatic summarization feature to condense lengthy answers to her interview questions for improved clarity.

Equipped with such AI-powered tools, Alice efficiently processes her remaining transcripts and utilizes our platform's analytical features to answer her research questions.

4 Benchmark of QDA-related NLP Tasks

This benchmark evaluates LLMs on tasks mirroring real-world QDA use cases, as outlined in the previous section. We aim to identify the most suitable model for effective user support to be integrated into DATS. To this end, we carefully select datasets for document classification (assigning tags to documents), document information extraction (extraction of metadata from documents), span classification (annotation of relevant passages), and text generation (correcting and summarizing text passages). Our platform mainly caters to English and German data, so we focus our evaluation and dataset selection on these two languages.

4.1 Models

Our open-source Discourse Analysis Tool Suite exclusively employs open-source and open-licensed libraries, which extends to integrated models. Since our primary users, universities and researchers, often handle sensitive data, local execution of the entire platform, including models, is crucial for maintaining data privacy. Given such environments' typically limited computational resources, we focus on small, efficient LLM variants for fast inference.

As a result, we evaluate three state-of-the-art open-source decoder-only models: Llama 3.1 (Dubey et al. 2024, Touvron et al. 2023), Gemma 2 (Gemma Team, 2024), and Mistral NeMo (Mistral AI Team 2024, Jiang et al. 2023) with 8B, 9B, and 12B parameters, respectively. This makes deploying them in environments with limited resources possible. We only test instruct fine-tuned models and use half-precision (FP16) variants. Llama 3.1 is an openly accessible, open-source model from Meta AI published under the Llama 3 Community License. It has a large context window of 128k tokens and was trained on a corpus of about 15 trillion multilingual tokens. Gemma 2 is a lightweight, open model from Google, built from the same technology as their Gemini models. This variant was trained on 8 trillion tokens from web documents, code, and mathematics, primarily in English. Mistral NeMo is a model from Mistral AI built in collaboration with NVIDIA and published under the Apache 2.0 license. It also offers a large context window of up to 128k tokens and was trained on multi-lingual and code data.

4.2 Experiment construction

We conduct zero-shot experiments using a single, clear prompt for each task, dataset, and language, deliberately avoiding extensive prompt engineering (prompts are detailed in appendix A, dataset taxonomies are listed in appendix B). Models are instructed on the expected output format, and deviations are counted as errors. We report formatting adherence at the end of this chapter.

All experiments ran on a single A100 GPU, repeated three times per configuration. Reported results are averaged across runs to mitigate fluctuations. We further aggregate results by model and task because of space restrictions; full results are listed in the appendix.

4.3 Task 1: Document classification

Document classification is essential in QDA for organizing data collections. Our platform supports fine-grained tag sets that enable researchers to classify documents into one or more groups. Many analysis features rely on these tags for comparison and sub-corpora creation, highlighting the potential of automatic tag suggestions to improve workflows.

We assess LLMs on two relevant document classification tasks. Multi-class involves assigning a single class from pre-defined options, further differentiated into coarse- and fine-grained settings based on the number of classes. Multi-label allows for multiple class associations.

4.3.1 Datasets

Tagesschau is an established German news website known for its serious and objective reporting covering news from Germany and the world. We automatically extracted a taxonomy of 4 main categories (coarse) and 20 sub-categories (fine) from a publicly available crawl⁴ spanning 2018 – 2023.

BBC is the public service broadcaster of the United Kingdom that publishes English news from UK and the world and is deemed a trusted source of information. Similarly, we extracted a taxonomy of 4 main categories (coarse) and 26 sub-categories (fine) from the BBC dataset published by Li et al. (2024) that covers 2017 – 2024.

IMDb, the Internet Movie Database, contains information about movies, TV shows, etc. The IMDb Genres dataset⁵ includes movie descriptions and their classification into one of 16 major genres (coarse) and 2-3 of 25 subgenres (multi-label). For all datasets, we sampled 10,000 documents.

Since analyzing news articles is a common use case within our Discourse Analysis Tool Suite, the Tagesschau and BBC datasets, which are news datasets, are a good fit for this benchmark. While not directly related to our domain, the IMDb Genres dataset was explicitly included for its multi-label classification task.

4.3.2 Results

We evaluate document classification on three sub-tasks: coarse-, fine-grained, and multi-label classification. The aggregated results are presented in Table 1, and the complete evaluation is in the appendix, Table 5. We report weighted Precision (Prec), Recall, F1, and Accuracy (Acc).

⁴ <https://github.com/bjoernpl/tagesschau>

⁵ <https://kaggle.com/datasets/rajucg/imdb-movies-dataset-based-on-genre>

Model	Task	Prec	Recall	F1	Acc
gemma2	coarse	67.20	65.42	64.85	65.4
llama3.1	coarse	64.33	59.36	58.07	59.4
mistral	coarse	64.39	62.00	61.28	62.0
gemma2	fine	70.50	62.00	63.00	62.4
llama3.1	fine	57.50	36.00	34.00	35.9
mistral	fine	66.50	53.50	55.00	53.6
gemma2	multi	55.48	52.23	52.35	8.6
llama3.1	multi	52.12	42.38	43.95	6.9
mistral	multi	57.51	49.62	51.05	8.6

Table 1: Evaluation of Task 1 - Document Classification

Gemma 2 consistently outperforms the other models, maintaining high performance (64.85 F1 coarse, 63.00 fine) despite the significant increase in classes for the fine-grained classification tasks. In contrast, the other models struggle in the fine-grained scenario, Llama 3.1 performing the worst (34.00 F1). We refer to appendix Table 5 to compare English and German performance. For coarse-grained classification, all models exhibit superior performance on the German Tagesschau dataset compared to the English BBC dataset but experience a noticeable drop (over 22 points F1) on the fine-grained German task. This suggests challenges in German fine-grained classification.

4.4 Task 2: Document Information Extraction

Similar to document classification, assigning metadata to research materials aids data organization in QDA. DATS enables users to define metadata that is used for features like search, filtering, visualization, and quantitative analysis. As automating metadata extraction could boost researcher productivity, this is a relevant task.

We frame it as document-level information extraction, exploring extractive question-answering (EQA) and template-filling approaches. EQA involves extracting answers from the context or labeling them as unanswerable. Template-filling (similar to slot-filling, relation extraction, and event extraction) focuses on extracting multiple related information, e.g., arguments of a relation or information about an event.

4.4.1 Datasets

The Stanford Question Answering Dataset (SQuAD) by Rajpurkar et al. (2016) is a benchmark dataset for extractive question-answering. It

consists of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a text segment from the context.

The SQuAD 2.0 dataset (Rajpurkar et al., 2018) builds upon the original one and introduces unanswerable questions, making it more challenging. Crowd workers carefully crafted the unanswerable questions to be similar to answerable ones.

The GermanQuAD dataset (Möller et al., 2021) is a German counterpart to SQuAD. This extractive question-answering dataset was carefully constructed by students and experts familiar with machine learning and QA on the German variants of the English Wikipedia articles used in SQuAD.

These datasets, with multiple questions per Wikipedia passage, align with our scenario of extracting various metadata from documents. While some questions align with typical metadata fields (e.g., "When," "How many," "Where"), others target more complex information, making them less suitable for metadata extraction. Nonetheless, strong performance on these datasets indicates potential for successful application in DATS.

The MUC-4 dataset (Sundheim, 1992), designed for template-filling, contains 1700 news articles about terrorist incidents. It requires systems to classify each incident and fill five slots of relevant information. This dataset aligns well with our use case of extracting multiple metadata from documents, particularly within the news domain, making it an ideal fit for our benchmark.

4.4.2 Results

We evaluate document-level information extraction on two sub-tasks: extractive QA and template-filling. The aggregated results are shown in Table 2, the complete evaluation in the appendix, Table 6. We report Exact Match (EM) and F1 scores, with template-filling scores averaged across all slots.

While Gemma 2 exhibits a clear advantage in extractive QA (79 F1), Llama 3.1 performs best in template-filling (40 F1). Interestingly, template-filling is a greater challenge overall, likely due to the increased complexity of extracting multiple correct answers simultaneously, as reflected in the lower scores across all models (at least 70 F1 for extractive qa vs. at most 40 F1 for template filling). In extractive QA, the models exhibit comparable performance on the GermanQuAD and SQuAD datasets (refer to appendix Table 6), indicating no significant difference between English and German language capabilities.

Model	Task	EM	F1
gemma2	extractive-qa	66.53	79.41
llama3.1	extractive-qa	56.04	70.64
mistral-nemo	extractive-qa	55.21	72.39
gemma2	template-filling	36.69	36.70
llama3.1	template-filling	40.62	40.63
mistral	template-filling	23.86	23.87

Table 2: Evaluation of Task 2 - Information Extraction.

4.5 Task 3: Span classification

Annotation (often also called coding) of relevant text passages (spans) is critical to many QDA projects and especially important for research projects following the Grounded Theory paradigm. Here, the coding is done in the three phases of "open," "selective," and "axial" coding. Our Discourse Analysis Tool Suite supports these coding phases. Automating parts of the annotation process could streamline their workflow, making this a relevant benchmark task. As users can create fine-grained code systems to annotate text passages in our platform, we formulate the automatic annotation of text passages as coarse- and fine-grained span classification tasks.

4.5.1 Datasets

Few-NERD (Ding et al., 2021) is a fine-grained, large-scale Named Entity Recognition dataset consisting of 8 coarse-grained and 66 fine-grained categories. Over 180,000 sentences of Wikipedia articles were carefully annotated by experienced annotators. German LER (Leitner et al., 2020), the German Legal Entity Recognition dataset, consists of German legal documents and a typology relevant to court decisions with 7 coarse-grained and 19 fine-grained types. About 66,000 sentences were annotated by two domain experts. This dataset is challenging, as models need to be familiar with German law terms. While categories like persons, events, and organizations are relevant, we often observe different annotations spanning multiple sentences in QDA projects. Still, FewNERD and German LER are included in our benchmark as NER is a prominent span classification task, and we argue that understanding the concepts of these datasets is likely required for performing well on more complex annotation tasks.

The dataset for quotation attribution in German news articles (Petersen-Frey and Biemann, 2024)

consists of 1000 annotated German news articles from WIKINEWS. It includes information about *who* said *what* to *whom* as well as *how* and *in which context*. Quotations are categorized into direct, indirect, free, and reported speech. This task is part of the benchmark, as the identification and annotation of utterances were common to multiple projects we conducted with colleagues from social sciences. However, we simplify this task by considering only "speaker" and "direct speech" annotations.

4.5.2 Results

We evaluate span classification on three sub-tasks: coarse-, fine-grained NER (coarse, fine), and quotations (quot). The aggregated results are listed in Table 3. The full evaluation is reported in the appendix, Table 7. We report weighted Precision (P), Recall (R), F1 score, and Accuracy (Acc). Accuracy includes the classification of outside tokens.

Gemma 2 consistently outperforms the other models across all sub-tasks (30–38 F1). Llama 3.1’s performance is notably poor (7–15 F1), primarily due to its frequent failure to adhere to output formatting instructions. The model often switches the positions of NER labels and corresponding text spans, leading to significant parsing errors. To maintain a fair comparison across all models, we retain our original parsing algorithm and prompt instead of fixing such errors. Consequently, our experiments indirectly evaluate the instruction following capabilities of the models. Increasing the number of classes from coarse to fine-grained NER leads to a significant performance drop across all models. Comparing English and German results (refer to appendix Table 7) reveals significantly lower scores for German. Even the best performing model, Gemma 2, achieves at most half the F1 score in German compared to English. This could suggest challenges in handling German text or reflect the increased difficulty of German LER.

4.6 Task 4: Text Generation

Researchers must potentially correct fluency issues or summarize long and wordy statements, especially when dealing with transcripts. Further, in the qualitative content analysis approach of Mayring (2019), popular in the Humanities in Germany, summarization, explication, and structuring are defined as the three main pillars of content analysis. Here, summarization is employed to condense information, explication to clarify meaning by providing context, and structuring to filter the material

Model	Task	P	R	F1	Acc
gemma2	coarse	35.19	44.07	37.98	85.02
llama3.1	coarse	26.22	4.70	7.82	80.33
mistral	coarse	39.28	15.38	21.22	81.76
gemma2	fine	40.25	31.64	31.84	83.86
llama3.1	fine	26.66	4.16	6.78	80.37
mistral	fine	39.72	10.66	15.39	81.04
gemma2	quot	36.40	25.61	29.81	90.29
llama3.1	quot	33.16	10.88	15.18	88.17
mistral	quot	23.03	10.69	13.71	86.24

Table 3: Evaluation of Task 3 - Span Classification

according to specific aspects systematically.

DATS allows users to attach notes to documents or text passages. As hinted at in Section 3, we aim to expand this functionality with fluency correction and text summarization capabilities. Assisting users with this can streamline their workflow. Consequently, we consider the two text generation tasks in our benchmark.

Abstractive summarization involves generating concise and fluent summaries similar to human-written ones and is thus likely preferred by our users. Disfluency correction aims to enhance readability by identifying and removing issues such as repetitions, filler words, and false starts.

4.6.1 Datasets

The Disfl-QA dataset (Gupta et al., 2021), initially intended for evaluating question-answering robustness against disfluencies, consists of about 12k pairs of fluent and corresponding disfluent questions built upon SQuAD 2.0. We repurpose the dataset to benchmark models’ ability to correct disfluent texts, leveraging only the question pairs.

The DISCO dataset (Bhat et al., 2023), designed to facilitate multilingual disfluency correction, comprises a human-annotated corpus of over 12k disfluent-fluent text utterance pairs in English, Hindi, German, and French. We utilize only the English and German parts. It is constructed upon a publicly available dataset of human-AI agent interactions. The dataset covers four disfluency types: Filler, Repetition, Correction, and False Start.

While the specific domains of these datasets (question-answering and human-AI interactions) differ from our platform’s use case of correcting transcriptions of interviews and other qualitative data, we included them due to the scarcity of re-

sources for disfluency correction.

The CNN/DM (Hermann et al., 2015) dataset, a widely-used benchmark for summarization tasks, consists of over 300,000 article-summary pairs extracted from CNN and Daily Mail news articles written in English between 2007 and 2015. The summaries are primarily based on human-generated highlights or article descriptions.

MLSUM (Scialom et al., 2020), a large-scale multilingual summarization dataset, comprises over 1.5 million article-summary pairs collected from online newspapers between 2010 and 2019. It covers five languages (French, German, Spanish, Russian, and Turkish) and is a multilingual extension to CNN/DM. We utilize the German part of MLSUM for our benchmark, drawn from the Süddeutsche Zeitung newspaper.

Both datasets, centered around news articles, align well with the use of news articles as a data source on our QDA platform, making them suitable for evaluating LLM summarization performance.

4.6.2 Results

We evaluate disfluency correction (CORR) and abstractive summarization (SUM). The aggregated results are shown in Table 4. The full evaluation is in appendix Table 8. We report Rouge (R), Exact Match (EM), F1 and METEOR scores.

Gemma 2 demonstrates superior performance in the disfluency correction task (84 R-1), with the other two models performing slightly worse (79 R-1). Llama 3.1 and Gemma 2 perform similarly in abstractive summarization (28 vs. 29 R-1). Both summarization and disfluency correction tasks exhibit a performance drop of at least 12 percentage points when applied to German text (see appendix Table 8), suggesting increased difficulty for all models in handling German text generation and instruction following.

4.7 Discussion

Across the four evaluated tasks, Gemma 2 consistently emerges as the top-performing model. While Llama 3.1 performs best in template-filling and abstractive summarization, its struggles with NER, instruction following and the consistent worst performance in German tasks highlight potential limitations. Mistral NeMo, while generally capable, could never demonstrate superior performance, even though it has the most parameters (12B) among the benchmarked models.

Regarding language performance, our findings reveal discrepancies. While all models exhibit strong performance on German data for coarse-grained document classification, a consistent decline is observed across the board for fine-grained document classification, span classification, and text generation tasks in German. This suggests that current open-source LLMs still face challenges in handling the complexities of the German language, particularly in nuanced and generative tasks.

Throughout all experiments, we tracked parsing errors. Gemma 2 consistently adhered to the provided instructions (1% issues), followed by Mistral NeMo (2% issues). Llama 3.1, on the other hand, struggled notably (5% issues), most evident by the span classification tasks.

Overall, our findings highlight the potential of LLMs, especially Gemma 2, to significantly enhance QDA workflows regarding document classification, document information extraction, span classification, and text generation.

5 Integrating LLMs in our QDA platform

User feedback during early testing led us to refine our envisioned workflow outlined in Section 3. The implementation of summarization and fluency correction aligned with our original plan. However, users wanted to access document-based assistance features (document tagging, metadata extraction, annotation) at any point in their workflow, not just during the initial document import. Consequently, we redesign the *LLM Assistant* as a standalone feature independent of the pre-processing pipeline.

The feature is initiated by selecting the documents for analysis and clicking the *LLM Assistant* button. This launches a five-step dialog. Step 5 is depicted in Figure 1, Step 1 - 4 are shown in appendix Figure 2:

1. Task Selection: Users select document tagging, metadata extraction, or annotation.
2. Focus: Users specify which tags, metadata fields, or codes the LLM should consider.
3. Prompt Review: Users revise the system- and user prompts that are automatically generated based on the chosen task and selections
4. Job Execution: A progress bar indicates that the LLM Assistant job runs in the background. The dialog can be closed and reopened later.
5. Result Review: LLM-generated suggestions and their reasoning alongside the documents'

Model	Task	Rouge 1	Rouge 2	Rouge L	Rouge Lsum	EM	F1	METEOR
gemma2	CORR	84.24	73.68	82.48	82.49	40.08	84.24	85.60
llama3.1	CORR	79.00	67.54	77.20	77.21	33.80	78.89	80.86
mistral	CORR	78.97	63.51	77.07	77.07	25.91	78.88	79.67
gemma2	SUM	27.89	8.86	18.96	21.92	0.00	26.90	23.68
llama3.1	SUM	29.06	9.93	19.56	22.70	0.00	28.15	25.70
mistral	SUM	24.81	7.86	17.04	19.54	0.00	24.35	21.82

Table 4: Evaluation of Task 4 - Text Generation

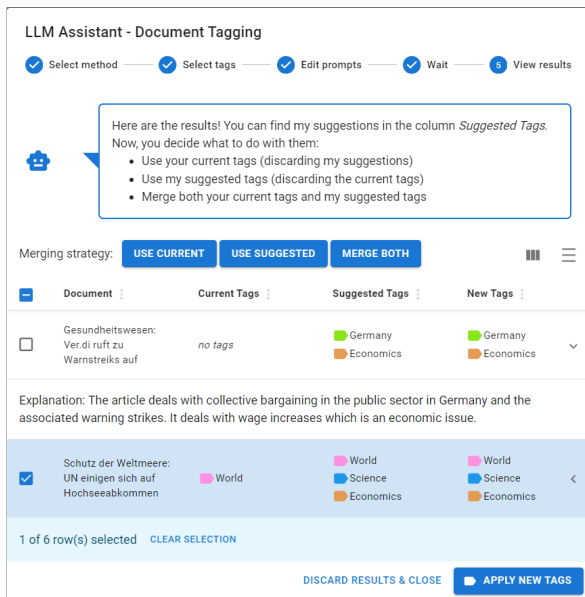


Figure 1: LLM Assistant - Step 5: Result View. Users can review the suggestions.

existing tags/metadata/codes are presented. Here, the user chooses to keep existing work, overwrite it with the LLM suggestions, or merge both per document or in batches.

User control and transparency are key requirements throughout the design of the *LLM Assistant*. Our philosophy is to ensure that any form of automation is a supportive tool, offering suggestions while the user retains decision-making. We firmly believe that AI should augment human expertise, not replace it. Thus, we’ve designed our platform to require explicit user approval for automatically-generated suggestions.

Our implementation achieves user control and transparency in two ways. The prompt review step allows users to inspect and modify the generated prompts, fostering transparency. Further, it enables advanced users to exert fine-grained control and provide additional task-specific instructions. In the

result view, we prioritize transparency by displaying the LLM’s reasoning and suggestions. Additionally, it allows users to critically evaluate the suggestions and decide how to incorporate them with their existing work, ensuring that the final output aligns with their intentions.

The *LLM Assistant* is built using React for the frontend and Ollama, FastAPI, and Celery for the backend. Celery handles background job processing, ensuring that LLM tasks run without interrupting user workflow. Ollama hosts the Gemma 2 model, which performed best in our benchmark. We reuse the benchmark prompts, as they were intentionally designed with future implementation in mind. We opt for the template-filling prompt style instead of extractive question-answering for metadata extraction. Providing metadata descriptions is more intuitive for users than formulating fitting questions for each field.

6 Conclusion

In this paper, we investigated the potential of LLMs to enhance qualitative data analysis workflows, focusing on common tasks within our open-source Discourse Analysis Tool Suite. We designed a benchmark reflecting real-world use cases and evaluated the performance of three prominent open-source LLMs. Our findings demonstrate the promise of LLM integration, particularly for English-language projects. Consequently, we implemented the *LLM Assistant* within our platform, a significant step towards empowering researchers with transparent and user-controlled AI assistance that augments, rather than replaces, human expertise.

In future work, we plan to extend the *LLM Assistant* to suggest new tags and codes, fostering a more exploratory QDA process. Furthermore, we aim to incorporate more domain-specific datasets that closely align with DH researchers’ real-world use

cases. Finally, we will explore few-shot learning approaches to enhance performance on nuanced tasks. User activities within our platform (e.g., tagging documents and annotating text passages) generate valuable training data for model fine-tuning, potentially leading to more efficient models tailored to individual user preferences. Code for replicating the benchmark⁶, the repository of DATS⁷ and a live demo are available⁸.

7 Limitations

While our implemented LLM Assistant demonstrates promising potential for enhancing QDA workflows, it's important to understand its limitations.

Firstly, the current implementation utilizes on zero-shot learning, which may not fully capture the nuances of specific QDA projects. Fine-tuning LLMs on user-specific data could lead to more accurate and contextually relevant suggestions.

Secondly, we restricted the LLM Assistant to English-language projects due to the observed performance discrepancies between English and German language tasks. Expanding language support will require further research and development to ensure similar performance across different languages.

Thirdly, the selection of suitable datasets for benchmarking remains a challenge. While we aimed to select datasets that closely resemble real-world QDA use cases, certain discrepancies between the benchmark tasks and actual user workflows exist. We will continue to identify and incorporate more representative datasets to ensure the evaluation's validity and generalizability.

Finally, we consider the inherent limitations of LLMs in general. They could struggle with tasks requiring complex reasoning, nuanced understanding of context, or common sense knowledge. Additionally, potential biases embedded within the training data can influence the LLM's outputs, requiring careful consideration and critical evaluation of the generated suggestions. This is why we deem the manual validation step integrated into our LLM Assistant, where users retain control over the acceptance and integration of AI-generated suggestions, as very important.

⁶ <https://github.com/uhh-It/llm4qda>

⁷ <https://github.com/uhh-It/dats>

⁸ <https://dats.ltdemos.informatik.uni-hamburg.de/>

References

- Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. *DISCO: A large scale human annotated corpus for disfluency correction in Indo-European languages*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12833–12857, Singapore. Association for Computational Linguistics.
- Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. *Few-NERD: A few-shot named entity recognition dataset*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3198–3213, Online. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. *The llama 3 herd of models*. *ArXiv*, abs/2407.21783.
- Gemma Team. 2024. *Gemma 2: Improving open language models at a practical size*. *ArXiv*, abs/2407.21783.
- Evelyn Gius, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, and Jan Horstmann. 2022. *CATMA: Computer Assisted Text Markup and Analysis*.
- Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. *Disfl-QA: A benchmark dataset for understanding disfluencies in question answering*. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3309–3319, Online. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, and Mantas Mazeika et al. 2021. *Measuring massive multitask language understanding*. In *Proceedings of the International Conference on Learning Representations*, Online.
- Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. *Teaching machines to read and comprehend*. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, page 1693–1701, Cambridge, MA, USA. MIT Press.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot et al. 2023. *Mistral 7b*. *ArXiv*, abs/2310.06825.
- Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. *A dataset of German legal documents for named entity recognition*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

- Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18600–18607, Vancouver, Canada.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, and Dilara Soylu et al. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*, 1525(1):140–146.
- Philipp Mayring. 2019. [Qualitative content analysis: Demarcation, varieties, developments](#). *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 20(3):Art. 16.
- Mistral AI Team. 2024. [Mistral nemo](#). <https://mistral.ai/news/mistral-nemo/>. Accessed: 2024-08-31.
- Timo Möller, Julian Risch, and Malte Pietsch. 2021. [GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval](#). In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Fynn Petersen-Frey and Chris Biemann. 2024. [Dataset of quotation attribution in German news articles](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 4412–4422, Torino, Italia. ELRA and ICCL.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *ArXiv*, abs/2212.04356.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, TX, USA. Association for Computational Linguistics.
- Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, and Wang Xiaofeng et al. 2024. [Can large language models serve as data analysts? a multi-agent assisted approach for qualitative data analysis](#). *ArXiv*, abs/2402.01386.
- Florian Schneider, Tim Fischer, Fynn Petersen-Frey, Isabel Eiser, Gertraud Koch, and Chris Biemann. 2023. [The D-WISE Tool Suite: Multi-Modal Machine-Learning-Powered Tools Supporting and Enhancing Digital Discourse Analysis](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 328–335.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. [MLSUM: The multilingual summarization corpus](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, and Abubakar Abid et al. 2023. [Beyond the imitation game: quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*, 2023(5):1–95.
- Anselm Strauss and Juliet Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. SAGE Publications, Inc.
- Anselm Strauss, Juliet Corbin, Solveigh Niewiarra, and Heiner Legewie. 1996. *Grounded Theory: Grundlagen Qualitativer Sozialforschung*. Beltz, Psychologie-Verlag-Union Weinheim.
- Beth M. Sundheim. 1992. [Overview of the fourth Message Understanding Evaluation and Conference](#). In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, McLean, VA, USA.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, and Marie-Anne Lachaux et al. 2023. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [Superglue: a stickier benchmark for general-purpose language understanding systems](#). In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, 50(1):237–291.

A Prompts

Table 9 lists all user prompts used in our experiments for English datasets and tasks. We translated system- and user prompts into German for German datasets. The system prompt is always the same:

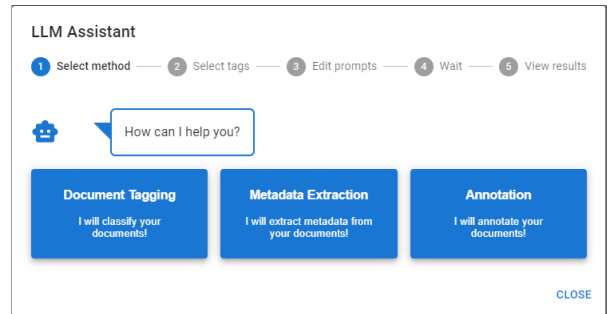
You are a system that supports the analysis of large amounts of text. You will always answer in the required format and use no formatting other than what the user expects!

All user prompts in this study adhere to a consistent structure designed to provide clear instructions and expectations. Each prompt begins with explicit task instructions, informing the model about the desired action, such as extracting an answer from a given context. This is followed by a detailed specification of the expected answer format, including potential responses for unanswerable questions (e.g., "Not answerable"). A concrete example is provided to clarify the desired output further. It is important to note that these examples are not few-shot examples derived from the datasets themselves. Finally, key constraints or limitations of the task are reiterated. We hope this ensures the model operates within the defined boundaries (e.g., extracting answers verbatim from the text).

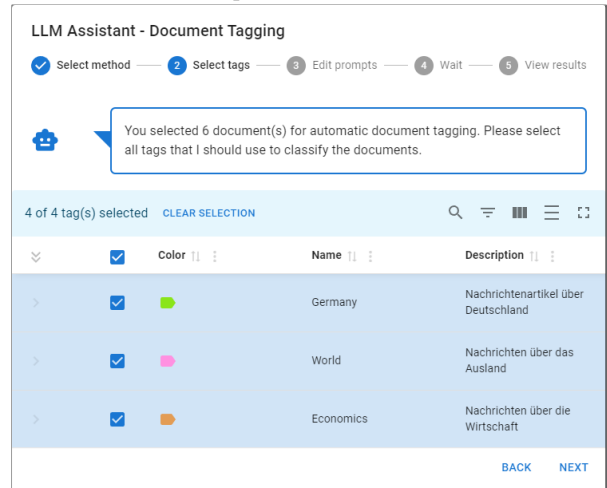
The placeholders ' {} ' within the prompt templates serve as dynamic variables populated with task-specific information. Depending on the task, these placeholders may contain a list of classes or categories for classification tasks, a set of slots for information extraction, or a specific question for question-answering tasks. Additionally, the context placeholder is filled with the relevant document or text passage from which the model is expected to derive its response.

B Additional Dataset Information

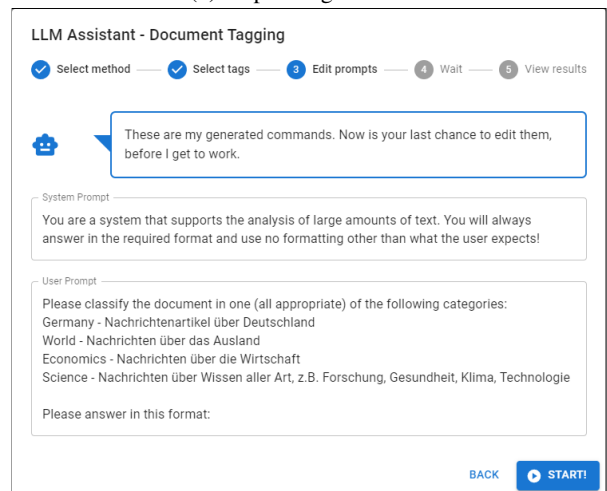
Table 10 provides additional information (e.g., taxonomies, slots) about the datasets used in our benchmark. All of this information was provided to the models within the prompt templates; for example, for the document classification task, the model was provided with a list of categories and their descriptions. For Task 1 - Text Classification, models were additionally provided with short 1-2 sentence descriptions of each class. Genre descriptions were taken from the IMDb website, news category descriptions were written by the authors



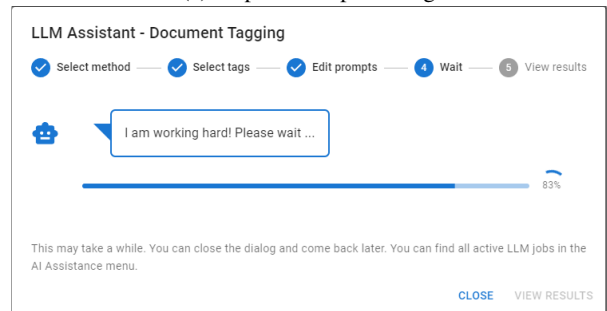
(a) Step 1: Method Selection



(b) Step 2: Tag Selection



(c) Step 3: Prompt Editing



(d) Step 4: Waiting

Figure 2: Steps 1-4 of the LLM Assistant Feature

Table 5: Full evaluation results of Task 1 - Document Classification

Model	Dataset	Language	Task	Precision	Recall	F1	Accuracy
gemma2	Tagesschau	de	coarse	84.88	81.94	82.01	81.94
llama3.1	Tagesschau	de	coarse	83.13	74.79	76.01	74.79
mistral-nemo	Tagesschau	de	coarse	83.55	77.62	78.92	77.62
gemma2	BBC	en	coarse	76.59	76.58	74.85	76.58
llama3.1	BBC	en	coarse	71.07	70.29	65.64	70.29
mistral-nemo	BBC	en	coarse	69.97	71.69	68.89	71.69
gemma2	imdb	en	coarse	40.12	37.74	37.70	37.74
llama3.1	imdb	en	coarse	38.78	33.00	32.57	33.00
mistral-nemo	imdb	en	coarse	39.66	36.70	36.04	36.70
gemma2	Tagesschau	de	fine	68.00	59.00	60.00	59.43
llama3.1	Tagesschau	de	fine	58.00	26.00	29.00	25.80
mistral-nemo	Tagesschau	de	fine	65.00	43.00	47.00	42.86
gemma2	BBC	en	fine	73.00	65.00	66.00	65.47
llama3.1	BBC	en	fine	57.00	46.00	39.00	46.07
mistral-nemo	BBC	en	fine	68.00	64.00	63.00	64.27
gemma2	imdb	en	multi-label	55.48	52.23	52.35	8.58
llama3.1	imdb	en	multi-label	52.12	42.38	43.95	6.88
mistral-nemo	imdb	en	multi-label	57.51	49.62	51.05	8.58

Table 6: Full evaluation results of Task 2 - Document Information Extraction

Model	Dataset	Language	Task	Exact Match	F1
gemma2	SQUAD1	en	extractive-qa	75.73	87.19
llama3.1	SQUAD1	en	extractive-qa	70.39	82.64
mistral	SQUAD1	en	extractive-qa	67.48	83.35
gemma2	SQUAD2	en	extractive-qa	63.85	69.80
llama3.1	SQUAD2	en	extractive-qa	48.82	56.25
mistral	SQUAD2	en	extractive-qa	50.81	60.64
gemma2	GermanQuAD	de	extractive-qa	60.01	81.25
llama3.1	GermanQuAD	de	extractive-qa	48.91	73.04
mistral	GermanQuAD	de	extractive-qa	47.34	73.18
gemma2	MUC4	en	template-filling	36.69	36.70
llama3.1	MUC4	en	template-filling	40.62	40.63
mistral	MUC4	en	template-filling	23.86	23.87

Table 7: Full evaluation results of Task 3 - Span Classification

Model	Dataset	Language	Task	Precision	Recall	F1	Accuracy
gemma2	fewnerd	en	coarse	48.53	55.72	51.15	83.73
llama3.1	fewnerd	en	coarse	40.6	7.72	12.87	79.46
mistral	fewnerd	en	coarse	49.95	18.97	26.32	81.12
gemma2	germanler	de	coarse	21.85	32.42	24.82	86.32
llama3.1	germanler	de	coarse	11.84	1.67	2.78	81.21
mistral	germanler	de	coarse	28.6	11.79	16.11	82.41
gemma2	fewnerd	en	fine	47.92	40.78	42.31	83.23
llama3.1	fewnerd	en	fine	38.97	6.53	10.54	79.2
mistral	fewnerd	en	fine	39.53	12.68	17.64	79.86
gemma2	germanler	de	fine	32.58	22.49	21.37	84.48
llama3.1	germanler	de	fine	14.36	1.78	3.02	81.54
mistral	germanler	de	fine	39.91	8.64	13.13	82.22
gemma2	quotations	de	quotations	36.4	25.61	29.81	90.29
llama3.1	quotations	de	quotations	33.16	10.88	15.18	88.17
mistral	quotations	de	quotations	23.03	10.69	13.71	86.24

Table 8: Full evaluation results of Task 4 - Text Generation. We report Rouge (R), Exact Match (EM), F1 and METEOR scores on two text generation tasks: disfluency correction (CORR) and summarization (SUM).

Model	Dataset	Lang	Task	R-1	R-2	R-L	R-Lsum	EM	F1	METEOR
gemma2	DisflQA	en	CORR	83.28	71.99	80.66	80.67	21.77	83.10	89.08
llama3.1	DisflQA	en	CORR	78.94	65.33	75.80	75.80	11.45	78.46	85.69
mistral	DisflQA	en	CORR	78.21	63.10	75.24	75.24	15.21	77.77	81.38
gemma2	DISCO	en	CORR	92.40	85.56	91.63	91.64	64.44	92.59	91.55
llama3.1	DISCO	en	CORR	91.20	84.57	90.58	90.58	64.01	91.43	91.07
mistral	DISCO	en	CORR	85.44	73.14	84.39	84.36	41.87	85.52	85.02
gemma2	DISCO	de	CORR	77.03	63.48	75.15	75.16	34.04	77.02	76.18
llama3.1	DISCO	de	CORR	66.87	52.73	65.21	65.24	25.94	66.77	65.83
mistral	DISCO	de	CORR	73.25	54.28	71.57	71.60	20.64	73.35	72.62
gemma2	CNNDM	en	SUM	34.98	11.21	22.51	28.43	0.00	33.30	29.73
llama3.1	CNNDM	en	SUM	36.44	12.44	23.09	29.38	0.00	34.88	31.97
mistral	CNNDM	en	SUM	30.81	9.44	20.49	25.48	0.00	30.19	24.01
gemma2	MLSUM	de	SUM	20.81	6.52	15.40	15.41	0.00	20.50	17.63
llama3.1	MLSUM	de	SUM	21.68	7.43	16.04	16.03	0.00	21.41	19.42
mistral	MLSUM	de	SUM	18.81	6.27	13.59	13.60	0.01	18.51	19.62

Table 9: The prompts used in both Evaluation and Implementation. {} are placeholders for task-dependent input.

Task	Prompt Template
Document Classification	<p>Please classify the document in one (all appropriate) of the following categories: {} Please answer in this format. You are not required to provide any reasoning. Category: <category> Reason: <reason> e.g. Category: News Document: {}</p>
Extractive QA	<p>Please extract the answer to the following question from the context below: Context: {} Question: {} Please answer in this format. If the question cannot be answered from the context, respond with 'Not answerable'. You are not required to provide any reasoning. Answer: <answer> or <not answerable> Reasoning: <reasoning> e.g. Answer: 42 Remember, the answer MUST be extracted verbatim from the text, do not generate it!</p>
Template Filling	<p>I prepared a list of slots. The slots are: {} Please extract the corresponding information (if any) from the following text: {} Please answer in this format. If the text does not include information about a specific slot, leave it empty. <Slot>: <extracted information> e.g. Incident: Arson, Perpetrator: John Doe, Weapon: Matches Remember, you MUST extract the information verbatim from the text, do not generate it!</p>
Summarization	<p>Please summarize the text below concisely, highlighting the most important information. Try to use about {} words only. Text: {} Respond in the following format: Summary: <summarized text> e.g. Summary: Theia was hit by a car ... Remember, you MUST summarize the original text, do not generate new facts!</p>
Disfluency Correction	<p>Please remove all disfluencies from the noisy, disfluent text below. Keep the text close to the original, but ensure it is read fluently. Text: {} Respond in the following format: Fluent text: <the corrected text> e.g. Fluent text: This picture looks great. Remember, you MUST keep to the original text; do not generate new content!</p>
Span Classification	<p>I prepared a list of categories/information. The categories are: {} Please extract fitting text spans from the following text: {} Respond in the following format: <category>: <extracted text> e.g. Art: Mona Lisa, Building: Eiffel Tower Remember, you MUST extract the information verbatim from the text, do not generate it!</p>

Table 10: Taxonomies of the datasets used in Task 1, 2, and 3.

Task 1	Document Classification
Dataset	Classes (coarse - fine)
BBC	UK - england, scotland, wales, ireland, politics World - africa, asia, australia, europe, latin, us, middle-east Sport - boxing, cricket, footbal, formula1, rugby, tennis Misc - business, education, elections, entertainment, arts, health, science, technology
Tagesschau	Inland - Deutschlandtrend, Gesellschaft, Innenpolitik, Mittendrin Ausland - Afrika, Amerika, Asien, Europa, Ozeanien Wirtschaft - Börse, Finanzen, Konjunktur, Technologie, Unternehmen, Verbraucher, Weltwirtschaft Wissen - Forschung, Gesundheit, Klima, Technologie
IMDB	Action, Adventure, Animation, Biography, Crime, Family, Fantasy, Film-Noir, History, Horror, Mystery, Romance, SciFi, Sports, Thriller, War
Task 2	Document Information Extraction
Dataset	Slots
MUC4	Incident: One of 'Arson', 'Attack', 'Bombing', 'Kidnapping', 'Robbery' Perpetrator: An individual perpetrator Group Perpetrator: A group or organizational perpetrator Victim: Sentient victims of the incident Target: Physical objects targeted by the incident Weapon: Weapons employed by the perpetrators
Task 3	Span Classification
Dataset	Classes (coarse - fine)
fewNERD	Art - broadcastprogram, film, music, other, painting, writtenart Building - airport, hospital, hotel, library, other, restaurant, sportsfacility, theater Event - attack, disaster, election, other, protest, sportsevent Location - GPE, bodiesofwater, island, mountain, other, park, road Organization - company, education, government, media, other, political party, religion, show organization, sportsleague, sportsteam Other - astronomy, award, biology, chemical, currency, disease, educational degree, god, language, law, living thing, medical Person - actor, artist, athlete, director, other, politician, scholar, soldier Product - airplane, car, food, game, other, ship, software, train, weapon
German-LER	Person - Anwalt, Richter Ort - Land, Stadt, Straße, Landschaft Organisation - Unternehmen, Institution, Gericht, Marke Norm - Gesetz, Verordnung, EU Norm Regulierung - Vorschrift, Vertrag Rechtsprechung Literatur
Quotations	Sprecher Direkte Rede

Cross-Dialectal Transfer and Zero-Shot Learning for Armenian Varieties: A Comparative Analysis of RNNs, Transformers and LLMs

Chahan Vidal-Gorène^{1,2}, Nadi Tomeh¹, Victoria Khurshudyan³

¹LIPN, CNRS UMR 7030, France

²École nationale des chartes, PSL University, Centre Jean Mabillon, France

³SeDyL, UMR8202, INALCO, CNRS, IRD, France

Correspondence: chahan.vidal-gorene@chartes.psl.eu

Abstract

This paper evaluates lemmatization, POS-tagging, and morphological analysis for four Armenian varieties: Classical Armenian, Modern Eastern Armenian, Modern Western Armenian, and the under-documented Getashen dialect. It compares traditional RNN models, multilingual models like mDeBERTa, and large language models (ChatGPT) using supervised, transfer learning, and zero/few-shot learning approaches. The study finds that RNN models are particularly strong in POS-tagging, while large language models demonstrate high adaptability, especially in handling previously unseen dialect variations. The research highlights the value of cross-variational and in-context learning for enhancing NLP performance in low-resource languages, offering crucial insights into model transferability and supporting the preservation of endangered dialects.

1 Introduction

This research is part of the DALiH project¹. The goal of the project is to create a robust digital platform for the exploration of both historical and contemporary forms of the Armenian language. The project intends to offer freely accessible and open-source resources, which will include grammatically annotated corpora and a variety of NLP tools such as annotation models, datasets, ASR models, among others.

This study focuses on four varieties of Armenian: Classical Armenian (CA), Modern Eastern Armenian (MEA), Modern Western Armenian (MWA), and the Getashen dialect (G). While CA, MEA, and MWA have annotated corpora and models included in the Universal Dependencies (UD) project, the

¹The DALiH project is funded by French National Research Agency ANR-21-CE38-0006.: Digitizing Armenian Linguistic Heritage: Armenian Multivariational Corpus and Data Processing

Getashen dialect, which is an underdocumented variety²

Linguistic resources like annotated corpora and NLP models for tasks such as lemmatization, POS-tagging, and morphological analysis are essential for both linguists and digital humanities scholars. These tools support tasks like corpus pre-annotation and the study of historical texts, oral traditions, and regional literature. In this study, we aim to develop models for these tasks across the four varieties, with a particular focus on the under-resourced Getashen dialect.

Our contributions are threefold:

1. Comparative Evaluation of Models: We conduct a comprehensive comparative study of traditional RNN models, state-of-the-art multilingual language models (e.g., mDeBERTa), and large models (e.g., ChatGPT) in zero and few-shot setups across the three UD-supported dialects.
2. Pre-annotation of the Getashen Dialect: We evaluate the feasibility of using existing NLP models to pre-annotate the Getashen dialect, assessing the adaptability and transferability of models trained on other dialects.
3. Variational divergences / convergences: We explore linguistic similarities between the dialects and how they affect model transferability, providing insights into cross-dialectal model adaptation.

2 Linguistic Background

Armenian is an Indo-European language that constitutes a distinct branch marked by both diachronic

²In our study, we employ two terms to characterize the status of linguistic data and resources: an "*underdocumented language*," which denotes a language or variety that lacks formal linguistic records, and an "*under-resourced language*," which signifies a language or variety that lacks digital resources.

and synchronic variation. The historical evolution of Armenian comprises: a) Classical Armenian (5th-10th centuries A.D.), b) Middle Armenian (11th-16th centuries), and c) Modern Armenian (17th century to the present). Synchronically, Modern Armenian includes Modern Eastern Armenian (MEA), Modern Western Armenian (MWA) and numerous Armenian dialects. All the written forms of the Armenian language utilize the unique Armenian alphabet.

While the linguistic divergences in historical variation are considerable, they vary in degree among the two modern standards and dialectal varieties, depending on the areal and genetic distance of each within the Armenian linguistic continuum (for more details, see [Sayeed and Vaux \(2017\)](#); for linguistic variation, see [Donabedian-Demopoulos \(2018\)](#) and [Khurshudyan and Donabédian \(2021\)](#)).

This study explores the morphological and morphosyntactic annotation of the following Armenian linguistic varieties: Classical Armenian, Modern Western Armenian, Modern Eastern Armenian and the Getashen dialect. Classical Armenian (CA) is typologically a right-branching language with complex inflectional morphology and syntactic constructions, featuring a tripartite nominative-accusative-genitive alignment and flexible word order. Beyond the Bible and other religious texts, numerous original and translated works across various genres (such as historiography, mathematics, medicine, etc.) have been preserved in Classical Armenian. Currently, it is maintained exclusively for liturgical purposes.

In contrast, MWA and MEA, and the dialect of Getashen are typologically left-branching, with nominative-accusative alignment and more agglutinative morphology. They exhibit a richer system of periphrastic constructions and maintain flexible word order. MWA and MEA were standardized in the mid-19th century, leading to a rich written heritage. Both standards are currently in use, with MWA primarily by the traditional Armenian diaspora, and MEA used in Armenia, Armenian community of Iran, and Armenian communities in ex-Soviet countries.

The Getashen dialect belongs to the -um branch of the Karabakh dialect family (for more details on Armenian dialects, see [Martirosyan \(2019\)](#) and [Davtyan \(1966\)](#)). This dialect is used in oral form. The data utilized in this study were recorded and transcribed as part of the project "Migration and Complex Identities in the Republic of Armenia (an

interdisciplinary study in anthropology and linguistics; Migrant Groups in Armenia (1940-2012): Parameters of Complex Identities)" during fieldworks conducted in Armenia in 2014-2015 ([Khurshudyan and Shagoyan, 2016](#)).

3 Related Work

Lemmatization, POS-tagging, and morphological analysis are fundamental tasks in NLP, historically tackled using RNN-based approaches ([Manjavacas et al., 2019](#)), or LSTM models often combined with pre-trained word embeddings like GloVe or FastText for contextual word representations. However, state-of-the-art transformer models like BERT ([Kondratyuk, 2019](#)), RoBERTa, and XLM-R have significantly outperformed these traditional methods by capturing long-range dependencies and contextual information more effectively through self-attention mechanisms, which allow them to manage complex linguistic patterns.

These transformer-based approaches, though highly effective, generally require large amounts of annotated data, making them less suitable for historical and under-resourced languages due to data scarcity. To address this challenge, recent trends have focused on leveraging attention mechanisms combined with transfer learning from models like DeBERTa ([Riemenschneider and Krahn, 2024](#)) or utilizing large language models (LLMs) in assisted annotation frameworks for these languages ([Zhao et al., 2024](#); [Bhat and Varma, 2023](#); [Kholodna et al., 2024](#)). Despite these advancements, very under-resourced languages like Coptic, Ancient Egyptian, or Old French still predominantly rely on Seq2Seq architectures, often using LSTM or GRU units with attention mechanisms to handle sequences and generate lemmas or morphological patterns ([Manjavacas et al., 2019](#); [Camps et al., 2021](#); [Sahala, 2024](#)).

In the case of Armenian, most lemmatization, POS-tagging, and morphological analysis experiments have focused on Modern Eastern Armenian ([Khurshudyan et al., 2022a](#); [Arkhangelskiy et al., 2012](#)) and Classical Armenian ([Vidal-Gorène and Kindt, 2020](#); [Kindt and Van Elverdinghe, 2022](#); [Kindt and Vidal-Gorène, 2022](#); [Kharatyan and Kocharov, 2024](#)), using LSTM, joint learning methods with RNNs or rule-based approaches ([Khurshudyan et al., 2022b](#)), achieving F1-scores ranging from 0.63 to 0.87 depending on the task and text genre (e.g., Gospel, colophon, HTR output or historiography). These methods have also been applied

	CA	MWA	MEA	G
Tokens	82,557	124,230	52,950	568
Unique tokens	6,837	27,773	14,320	377
Unique lemma	2,472	11,952	7,087	248
Sentences	4,146	6,656	2,500	100
Sentence length (min/max/mean)	2 / 97 / 19.91	1 / 189 / 18.66	2 / 121 / 21.18	27 / 98 / 56.8
Word length (min/max/mean)	1 / 17 / 3.48	1 / 37 / 4.97	1 / 48 / 4.97	1 / 13 / 4.67

Table 1: Overview of the four datasets, including the total number of tokens, unique tokens, unique lemmas, number of sentences, and distributions of sentence and word lengths in defined subsets.

to MEA with similar results (Vidal-Gorène et al., 2020). Experiments in transferring MEA models to Armenian dialects, including MWA, have reported accuracies around 65% in lemmatization and 80% in POS-tagging (Vidal-Gorène et al., 2020).

The application of transformer models or LLMs to Armenian linguistic tasks remains in its early stages, with current usage primarily in classification tasks (Avetisyan et al., 2023).

4 Armenian Datasets

This study draws upon four datasets representing different Armenian dialects. Three of these datasets are sourced from the Universal Dependencies (UD) project (de Marneffe et al., 2021): Classical Armenian (CA)³, Modern Eastern Armenian (MEA)⁴, and Modern Western Armenian (MWA)⁵. The fourth dataset, representing the Getashen (G) dialect, was compiled and transcribed as part of the project "Migration and Complex Identities in the Republic of Armenia" (Khurshudyan and Shagoyan, 2016).

The UD datasets are designed to provide comprehensive morphological and syntactic annotations following UD guidelines, covering a wide range of Armenian language varieties. In contrast, the Getashen dataset consists of raw transcribed text, from which a small number of sentences have been manually annotated specifically for this study.

Modern Eastern Armenian The MWA dataset, also developed by the ArmTDP team, comprises around 52,950 tokens in 2,500 sentences. It spans a wide variety of genres, including blogs, fiction, legal texts, and news. Each sentence is annotated with lemmas, Universal POS-tags (UPOS), and various morphological features, making it the largest manually verified corpus of Eastern Armenian,

complete with detailed dependency trees for every sentence.

Modern Western Armenian The MWA dataset, developed by the ArmTDP team, is the most extensive among them, featuring approximately 124,230 tokens across 6,656 sentences, covering a broad range of genres such as blogs, fiction, and nonfiction. The annotation process mirrors that of the MEA dataset, combining automatic glossary-based scripting with manual revision. This dataset is the only manually verified corpus of Western Armenian, offering comprehensive morphological and syntactic annotations.

Classical Armenian The CA dataset is a treebank of the Classical Armenian translation of the four Gospels, by the Classical Armenian-CAVaL treebank project, containing 82,557 tokens in 4,146 sentences. Initially annotated in a non-UD style as part of the PROIEL project, it was later converted to UD format through a rule-based process, followed by manual corrections to ensure accuracy.

Getashen Armenian The fourth dataset, representing the Getashen (G) dialect, consists of a smaller collection of 100 manually annotated sentences. It is used to investigate the transferability of models trained on well-established language variants with long-standing writing traditions and consistent annotation schemas (such as the UD datasets) to a less-documented dialect.

Dataset Statistics Table 1 provides a detailed overview of the composition of these datasets, including statistics on tokens, unique tokens, lemmas, sentences, and the length distributions of both sentences and words. The MEA and MWA datasets, being the largest, show complete alignment in POS-tags, indicating that all POS-tags present in one are also found in the other. They also share the highest overlap in tokens (7.90%) and lemmas (14.25%),

³https://universaldependencies.org/treebanks/xcl_caval/

⁴https://universaldependencies.org/treebanks/hy_armtdp/

⁵https://universaldependencies.org/treebanks/hyw_armtdp/

	Tokens	Lemmas	POS
MEA-MWA	7.90	14.25	100.00
MEA-CA	2.65	4.77	94.44
MWA-CA	3.31	6.21	94.44
G-MWA	32.36	43.54	36.36
G-MEA	21.22	33.06	36.36
G-CA	11.67	13.70	36.36

Table 2: Percentage overlap (intersection/union) of unique tokens, lemmas, and POS-tags between the four dialect datasets.

suggesting a relatively high degree of linguistic similarity between these two dialects. Table 2 further elaborates on these commonalities, showing that while the MEA-MWA pair exhibits the greatest overlap, the MEA-CA and MWA-CA comparisons have lower overlap in both tokens (2.65% and 3.31%, respectively) and lemmas (4.77% and 6.21%, respectively). This suggests a more distinct linguistic boundary between these datasets.

The Getashen (G) dataset, consisting of transcriptions of spoken language, shows a relatively low overlap with other datasets, ranging from 11.67% to 32.36% for tokens and 13.70% to 43.54% for lemmas. The low overlap in both tokens and lemmas likely reflects the differences inherent in transcriptions of spontaneous speech compared to written text, where greater variability and a broader vocabulary are common. Additionally, the Getashen dataset has an unusually high mean sentence length of 56.8 tokens, contrasting with the shorter averages in the other datasets, which may underscore the complexity and fluidity of spoken language as compared to more structured written forms.

5 Methodology

Our approach aims to understand how different models perform on token-level annotation tasks — lemmatization, POS-tagging, and morphological feature tagging — across multiple Armenian dialects with varying levels of resources and label sets. We explore a unified sequence labeling framework to handle these tasks, leveraging different model architectures, including RNNs, pre-trained transformers (mDeBERTa), and large language models (LLMs). By comparing these models in supervised, transfer learning, and zero/few-shot learning settings, we study how well they generalize across dialectal variations and whether com-

binning data from multiple dialects improves performance, particularly for those with limited training data. Codes and raw results are available on Github.⁶

5.1 Task Modeling

The tasks considered in this study — lemmatization, POS-tagging, and morphological feature tagging — are all treated as sequence labeling problems. For each task, a sequence of words (tokens) in a given sentence is mapped to a sequence of labels. *Lemmatization* involves mapping each token to its dictionary form, *POS-tagging* assigns each token its corresponding part-of-speech tag, and *morphological tagging* annotates each token with relevant morphological features (such as case, person, and number).

5.2 Model Architectures

We compare three types of model architectures for the sequence labeling tasks:

Recurrent Neural Network (RNN): An RNN specialized for linguistic tasks (Vidal-Gorène and Kindt, 2020), which builds on has already been used for CA (Vidal-Gorène and Kindt, 2020) and MEA (Vidal-Gorène et al., 2020). Our model relies on the PIE architecture (Manjavacas et al., 2019). This method improves annotation of non-standard languages by using an encoder-decoder architecture based on Recurrent Neural Networks (RNNs), enriched with sentence context information through a hierarchical bidirectional RNN and a joint learning approach with a bidirectional language modeling loss. We slightly modify the architecture, adding an attention layer. The RNN models for lemmas, POS-tags and for each morphological feature are trained separately since our preliminary experiments showed that joint training did not help.

Pretrained Bi-Encoder Transformer: A pre-trained mDeBERTa model (He et al., 2021), a multilingual variant of the DeBERTa model, finetuned on the dataset of each dialect. This architecture leverages the power of transformer-based contextual embeddings. Each model consists of the mDeBERTa model, followed by a dropout layer and a linear classifier. Using this setup, the hidden states from the mDeBERTa transformer are mapped to logits that correspond to the labels of each of the

⁶<https://github.com/CVidalG/dalih-corpora/>

tasks. Models for all tasks share the same backbone transformer and differ only in the classification heads.

Large Language Model (LLM): We employ ChatGPT-4 (OpenAI, 2024) a pretrained large language model in zero-shot and few-shot settings to evaluate its ability to perform the sequence labeling tasks across dialects.

5.3 Learning Paradigms

We explore multiple data setups and learning paradigms to evaluate model performance across different scenarios:

In-Domain Supervised Learning: Each model (RNN and mDeBERTa) is trained in a supervised manner on a specific dialect and evaluated on the same dialect to establish baseline performance.

Cross-Dialect Transfer Learning: To assess the transferability of knowledge, models trained on one dialect are directly evaluated on other dialects without any adaptation. This setup helps us understand how well the models generalize across dialects with different label sets and linguistic characteristics.

Multi-Dialect Supervised Learning: We train the models on the combined datasets of all four varieties to see if pooling data improves performance, especially for dialects with limited training data.

Zero and Few-Shot Learning: We only used ChatGPT-4 in this setup. We aim to evaluate the ability of LLMs to generalize across dialects without explicit training on each. In the few-shot setup, ChatGPT was exposed to a small number of labeled examples using In-Context Learning (ICL) (Brown et al., 2020). We employed three sampling strategies for generation: sequence sampling, random sampling, and less frequent sampling, the latter two strategies performing well in annotation tasks (Bansal and Sharma, 2023). We used sample sizes of 10, 50, 100, and 500 tokens. Experiments were repeated three times, and results were averaged.

5.4 Evaluation Metrics

We use the macro-averaged F1-score instead of the micro-average to give equal weight to all classes, ensuring that the performance on less frequent classes is fairly represented.

	CA	MWA	MEA	G
Lemma				
RNN	0.66	0.91	0.79	-
mDeBERTa	0.70	0.44	0.36	-
LLM zero-shot	0.62	0.83	0.74	0.83
LLM in-context	0.74	0.83	0.83	-
POS				
RNN	0.98	0.98	0.98	-
mDeBERTa	0.91	0.90	0.88	-
LLM zero-shot	0.87	0.86	0.91	0.86
LLM in-context	0.91	0.91	0.85	-
Features				
RNN	0.88	0.70	0.66	-
mDeBERTa	0.88	0.78	0.77	-
LLM zero-shot	0.84	0.71	0.81	-
LLM in-context	0.86	0.75	0.88	-

Table 3: F1 macro average results for in-domain supervised learning. The G dialect does not make use of the UD system for features and is not evaluated.

6 Results

6.1 Main Results: Overall Comparison

The results presented in Table 3 show that the RNN consistently performs well across all tasks, particularly for POS-tagging, where it achieves near-perfect scores across the dialects. However, the LLM in-context method often matches or outperforms the RNN for lemmatization and morphological feature tagging, especially in the MEA, indicating its strong adaptability and context understanding. Interestingly, mDeBERTa lags behind in several tasks, particularly for lemmatization in the MEA dialect, suggesting that fine-tuning pretrained models may not always be advantageous compared to both RNNs (specifically designed for the task) and ChatGPT-4. ChatGPT’s performance in zero-shot setups also shows its potential for generalization, especially for the G dialect where it performs comparably to supervised methods.

6.2 In-Domain Supervised Learning

We further analyzed the performance of the RNN model which demonstrates strong performance for both lemmatization and POS-tagging on known tokens, achieving high F1-scores across the dialects (e.g., 0.94 for MWA in lemmatization and 0.99 for MWA in POS-tagging). However, its performance significantly drops on unknown tokens, with F1-

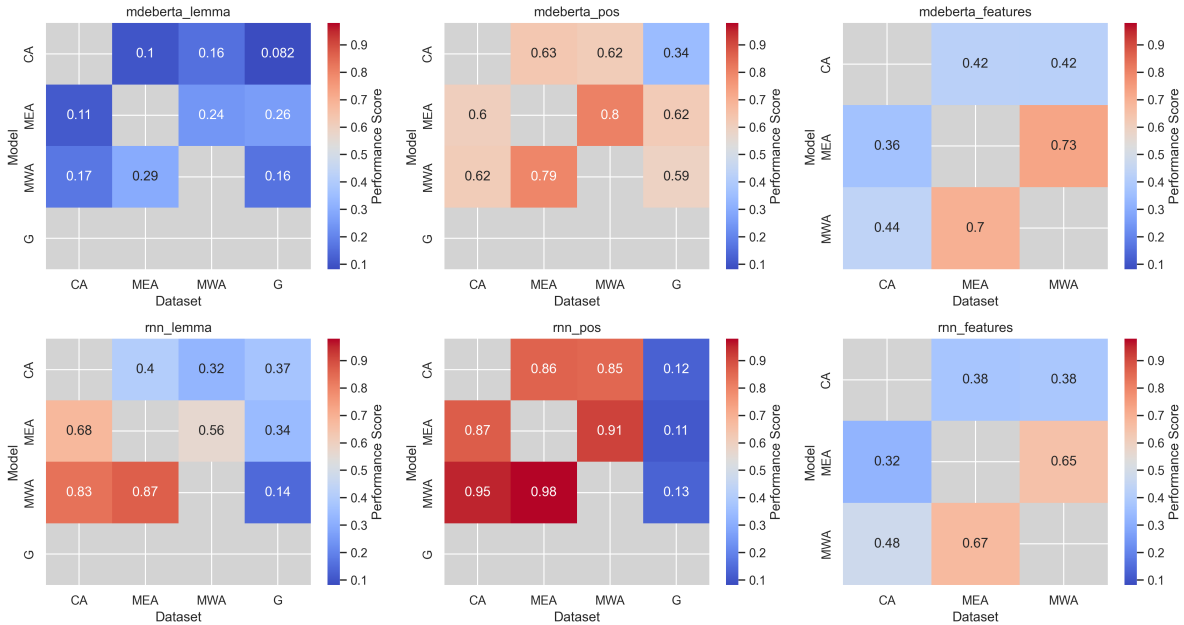


Figure 1: Cross-dialect performance of mDeBERTa and RNN models: Each model was evaluated on datasets outside of its training dialect to assess the generalization capability and immediate applicability without dialect-specific adaptation or mapping. Note that no model is trained on the G dialect as we only have a test set. This G test set contains only lemma and POS annotations.

scores decreasing to 0.43 for MEA and 0.53 for CA in lemmatization, and to 0.50 for CA and 0.53 for MEA in POS-tagging. These results indicate that while the RNN model is effective for known tokens, it struggles with less frequent or unseen classes, particularly in the lemmatization of MEA and CA. In comparison, mDeBERTa does not seem to suffer from this problem, which suggests that combining both models would be beneficial.

While the RNN and mDeBERTa models achieve similar overall performance when averaged across all features, a closer examination reveals that both models excel at handling certain morphological features, such as deixis and tense in MEA, and polarity and person in MWA, with F1-scores near or at 1.0. However, they perform poorly on features related to politeness, degree, and stylistic variations, suggesting that the models are particularly challenged by features that are less frequent or more nuanced in their expression.

6.3 Cross-Variational Transfer Learning

Performance Across Armenian Variation The comparison between mDeBERTa and RNN models across the Armenian dialects (CA, MEA, MWA) highlights the potential and challenges of cross-dialectal modeling for low-resource languages (Figure 1). The RNN consistently outperforms mDe-

BERTa in lemmatization and POS-tagging, with lemmatization scores ranging from 0.32 to 0.87 and POS-tagging scores from 0.85 to 0.98, compared to mDeBERTa’s lower range (0.10 to 0.29 for lemmatization and 0.60 to 0.80 for POS-tagging). However, mDeBERTa performs better on morphological features, achieving scores from 0.36 to 0.73, implying a capacity to handle more generalized tasks despite not being specifically tailored for them.

Dialect Compatibility The results indicate strong compatibility between MEA and MWA for both lemmatization and POS-tagging, reflecting their shared morphological and syntactic structures, with the highest cross-dialect scores at 0.87 and 0.98, respectively. Conversely, the CA dialect shows lower compatibility with modern dialects, particularly in transferring morphological features, where the best CA-to-MWA score is 0.44, pointing to significant linguistic divergence.

Generalization to New Dialects For the new dialect G, although neither model has been specifically trained on it, mDeBERTa and the RNN demonstrate reasonable performance, particularly in POS-tagging and lemmatization (best scores of 0.62 and 0.37, respectively). These findings suggest that cross-lingual transfer and general-purpose models can be valuable for handling linguistic tasks

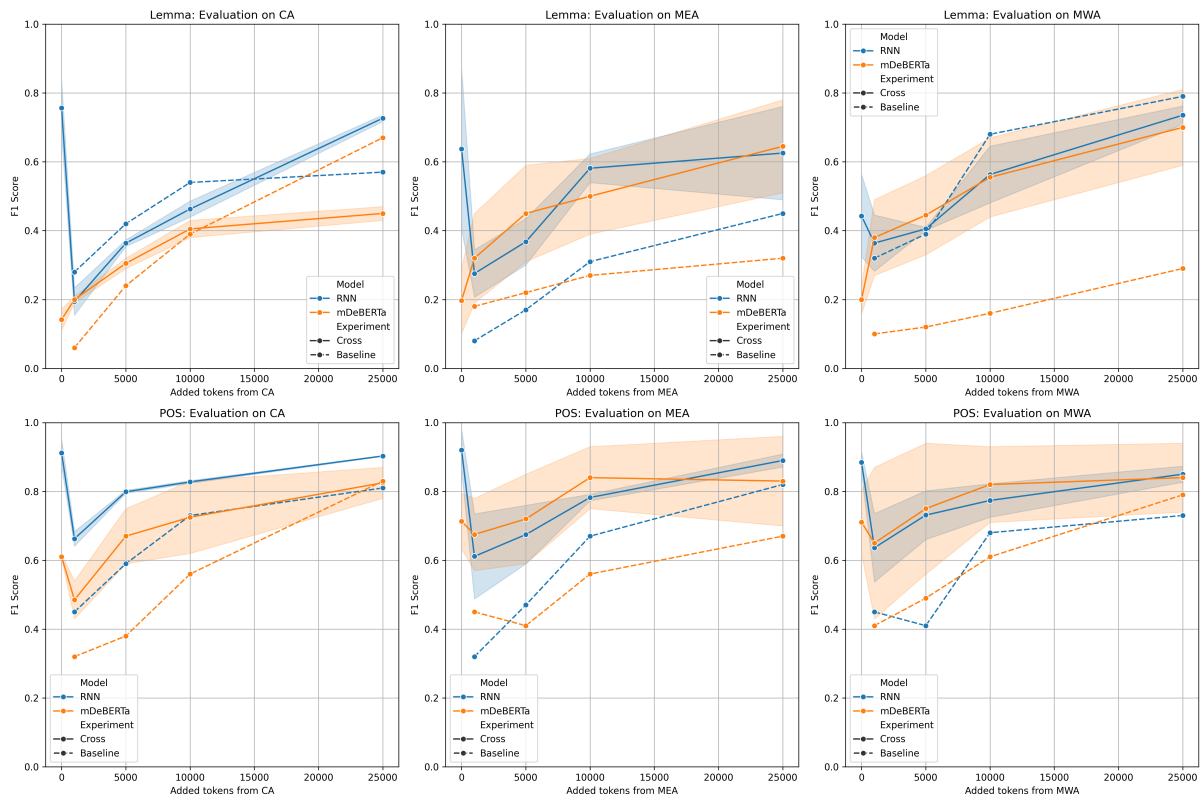


Figure 2: Performance comparison of lemmatization and POS-tagging in multi-dialect supervised learning versus zero-shot transfer learning. Error-bars represent the range of F1-score depending of the used base model (e.g. MWA + CA and MEA + CA for the first subplot)

in under-resourced languages, providing a practical alternative to task-specific models when extensive training data is unavailable.

6.4 Multi-Dialect Supervised Learning

Figure 1 illustrates that adding data from a target dialect to a model trained on a different dialect (“non-specialized model“) does not always improve performance. For instance, a non-specialized RNN trained on MEA and evaluated on CA initially achieves a strong F1 score of around 0.68. However, as CA data is incrementally added, the performance of this RNN decreases — dropping to 0.15 with just 1,000 CA tokens. Only after adding 25,000 tokens does the RNN’s performance recovers to an F1 score of approximately 0.74, aligning with its initial performance.

In contrast, mDeBERTa, which also starts as a non-specialized model with an F1-score of 0.11 on CA without any CA data, benefits more from adding targeted CA data. By incorporating 10,000 CA tokens, its F1 score rises to 0.43, and with 25,000 tokens, it reaches 0.67, nearly matching the performance of the RNN.

Interestingly, across all evaluation sets (CA, MEA, MWA), non-specialized models (those trained on one dialect and tested on another) often outperform specialized models (those pre-trained and fine-tuned by adding data from the same dialect as the evaluation set). For example, the non-specialized RNN evaluated on MEA without any added MEA data outperforms the specialized RNNs trained directly on MEA, until a significant amount of MEA data is added. This finding highlights the effectiveness of a cross-dialect approach, where training on data from different dialects can lead to better generalization than focusing solely on the target dialect.

6.5 LLM with Few and Zero-Shot Learning

Our goal was to assess how sampling strategy and sample size affect model performance in lemmatization, POS-tagging and full morphological analysis. Evaluations were conducted on a 200-token subset from the test dataset of each language, representing zero-shot performance and varying levels of in-context learning. Results are summarized in Figure 3.

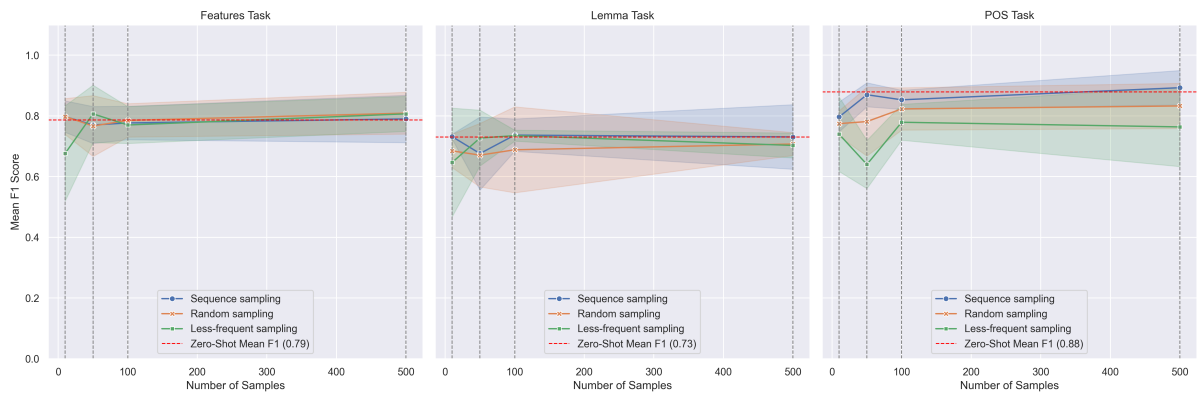


Figure 3: In-context learning using ChatGPT4 with three different sampling strategies: evolution of the mean F1-macro across CA, MEA and MWA.

Overall, the results demonstrate that in-context learning is particularly effective for lemmatization, with sequence sampling consistently outperforming other methods across all tasks and languages. However, for POS-tagging and morphological analysis, the LLM already achieves good results in the zero-shot setting, even for a very under-resourced dialect like G, and further improvements through in-context learning are less pronounced. Sequence sampling works better because it preserves the natural context of tokens, allowing the model to better understand and predict the linguistic patterns within the data. In contrast, random and less frequent sampling disrupts this context, leading to more variability and instability in the results.

Lemmatization The zero-shot F1 score for CA was 0.62, improving to 0.96 with just 10 samples using sequence sampling. For MEA, the zero-shot F1 score was 0.51, which improved significantly with 500 samples, achieving an F1 of 1.0. MWA started with a zero-shot F1 of 0.79, achieving 0.86 with sequence sampling, though additional samples did not consistently lead to improvements. Zero-shot and in-context F1 score on G is similar to MWA.

POS-Tagging POS-tagging using ChatGPT-4 began with a zero-shot F1 score of 0.87 for CA, which remained robust across all sampling strategies. For MEA, the performance improved steadily with sample size, especially with sequence sampling. MWA showed strong initial performance, but additional samples did not yield substantial improvements. Zero-shot and in-context F1 score on G is also similar to MWA.

Morphological Analysis In morphological analysis, sequence sampling led to stable and high F1 scores across all languages, though less frequent sampling exhibited more variability. For CA and MEA, sequence sampling consistently outperformed other methods, while improvements plateaued quickly for MWA.

7 Conclusion

The results from this study demonstrate the varying strengths of different model architectures in handling token-level annotation tasks across multiple Armenian dialects. RNN models consistently performed well, particularly in POS-tagging, where they achieved near-perfect F1 scores of up to 0.98, showcasing their robustness for tasks with known tokens. However, they struggled with less frequent or unseen tokens, where the adaptability of large language models (LLMs), especially in zero-shot and few-shot learning setups, became evident. For instance, ChatGPT-4 in zero-shot learning achieved an F1 score of 0.83 in lemmatization for the Getashen dialect. Pretrained transformers like mDeBERTa, while showing potential in handling morphological features with F1 scores reaching 0.73, often lagged behind RNNs and LLMs, particularly in lemmatization tasks, where their performance dropped to as low as 0.36 in the MEA dialect. Cross-dialect transfer learning revealed that non-specialized models can often generalize better across dialects than specialized ones, suggesting that a cross-dialect approach may be more effective for low-resource languages. In-context learning with LLMs further highlighted their ability to rapidly adapt and improve performance, particularly in lemmatization, where sequence sampling

led to an increase in F1 scores from 0.62 to 0.96 with just 10 samples. In the future, we plan to generalize our approach to include multiple other dialects and to ensure normalization of annotations, facilitating more consistent and comprehensive linguistic analysis.

Perspectives on Interoperability optimization

The annotations applied to the Armenian linguistic data exhibit variability across several dimensions. Firstly, the linguistic level encompasses various types of annotation, including morphological annotation, which involves part-of-speech tagging and the specification of full morphological features, as well as syntactic, semantic, and lexical annotations. Additionally, different categories are utilized to classify various linguistic phenomena, and distinct tagging systems are employed for different features within each annotation type. Moreover, there are notable differences in how morpheme glossing is split. The principles of tokenization are also significant, encompassing considerations such as the presence or absence of spaces and the treatment of internal and external punctuation marks. Finally, the diversity of target language varieties further influences the annotation process.

A potential avenue for further advancement could involve establishing tagging alignment and normalization among the existing datasets. However, automatic mapping without prior analysis and matching is not feasible, as the datasets employ different tagging principles, even though three of them are annotated within the Universal Dependencies framework. A significant systematic issue arises with the MWA and MEA datasets, where formal and functional criteria are mixed, whereas the Classical Armenian dataset employs exclusively formal criteria for tagging.

Another possible approach for dialectal data, for which no written tradition exists, is to process this data through normalization with either the MWA or MEA datasets (Arkhangelskiy and Georgieva, 2018; Waldenfels von R. and Dobrushina, 2014). While this approach may be beneficial for dialectal data, it also necessitates preliminary analysis and specific mapping.

While the aforementioned approaches can lead to significant improvements, establishing a fully harmonized and interoperable annotation system across all projects remains unattainable due to vari-

ations in project objectives, linguistic preferences, and the contextual conditions under which these systems were developed. Nevertheless, two parallel pathways can be explored: first, analyzing the existing systems to propose conversion options between them; and second, formulating common principles and annotation solutions for Armenian language data that could be embraced by the user community, while also allowing for conversion into various annotation systems as needed.

8 Acknowledgement

The DALiH project is funded by French National Research Agency ANR-21-CE38-0006.

References

- Timofey Arkhangelskiy, Oleg Belyaev, and Arseniy Vydrin. 2012. The creation of large-scale annotated corpora of minority languages using uniparser and the eanc platform. In *Proceedings of COLING 2012: Posters*, pages 83–92.
- Timofey Arkhangelskiy and Ekaterina Georgieva. 2018. Sound-aligned corpus of udmurt dialectal texts. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 26–38. Association for Computational Linguistics.
- Karen Avetisyan, Arthur Malajyan, Tsolak Ghukasyan, and Arutyun Avetisyan. 2023. A simple and effective method of cross-lingual plagiarism detection. *arXiv preprint arXiv:2304.01352*.
- Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.
- Savita Bhat and Vasudeva Varma. 2023. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 100–107.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

- Jean-Baptiste Camps, Thibault Clérice, Frédéric Duval, Naomi Kanaoka, Ariane Pinche, et al. 2021. Corpus and models for lemmatisation and pos-tagging of old french. *arXiv preprint arXiv:2109.11442*.
- Karo Davtyan. 1966. *Lernayin Ġarabaġi barbarayin k'artezəf= The dialectal map of Nagorno-Karabakh*. Yerevan.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. **Universal Dependencies**. *Computational Linguistics*, 47(2):255–308.
- Anaid Donabedian-Demopoulos. 2018. **Middle East and Beyond - Western Armenian at the crossroads : A sociolinguistic and typological sketch**. In Christiane Bulut, editor, *A sociolinguistic and typological sketch, in Bulut, Christiane, Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery, , 111/2018, Harrazowitz Verlag*, volume 111 of *Turcologica*, pages 89–148. Harrazowitz Verlag.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. **{DeBERTa}: {Decoding}-{enhanced} {bert} {with} {disentangled} {attention}**. In *International Conference on Learning Representations*.
- Lilit Kharatyan and Petr Kocharov. 2024. Development of Linguistic Annotation Toolkit for Classical Armenian in SpaCy, Stanza, and UDPipe. In *Proceeding of The First Workshop on Data-driven Approaches to Ancient Languages (DAAL 2024)*, pages 58–66.
- Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. *arXiv preprint arXiv:2404.02261*.
- Victoria Khurshudyan, Timofey Arkhangelskiy, Misha Daniel, Vladimir Plungian, Dmitri Levonian, Alex Polyakov, and Sergei Rubakov. 2022a. **Eastern Armenian national corpus: State of the art and perspectives**. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 28–37, Marseille, France. European Language Resources Association.
- Victoria Khurshudyan, Timofey Arkhangelskiy, Misha Daniel, Vladimir Plungian, Dmitri Levonian, Alex Polyakov, and Sergei Rubakov. 2022b. Eastern armenian national corpus: State of the art and perspectives. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 28–37.
- Victoria Khurshudyan and Anaïd Donabédian. 2021. **Cleft constructions and focus strategies in modern armenian**. *Faits de Langues*, 52(1):89 – 116.
- Victoria Khurshudyan and Gayane Shagoyan. 2016. Obscured borders of migrants’ ‘locality’: Language and identity shift of armenian refugees from azerbaijan: Case study of getashen village. In *Language Indexicality and Belonging Conference*.
- Bastien Kindt and Emmanuel Van Elverdinghe. 2022. Describing language variation in the colophons of armenian manuscripts. In *Proceedings of the Workshop on Processing Language Variation: Digital Armenian (DigitAm) within the 13th Language Resources and Evaluation Conference*, pages 20–27.
- Bastien Kindt and Chahan Vidal-Gorène. 2022. From manuscript to tagged corpora. *Armenica-International Journal of Armenian Studies*, 1:73–96.
- Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual bert fine-tuning. In *Proceedings of the 16th workshop on computational research in phonetics, phonology, and morphology*, pages 12–18.
- Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*.
- Hrach Martirosyan. 2019. **2.2. The Armenian dialects**, pages 46–105. De Gruyter Mouton, Berlin, Boston.
- OpenAI. 2024. Chatgpt-4. <https://openai.com>.
- Frederick Riemenschneider and Kevin Krahn. 2024. Heidelberg-boston@ sigtyp 2024 shared task: Enhancing low-resource language analysis with character-aware hierarchical transformers. *arXiv preprint arXiv:2405.20145*.
- Aleksi Sahala. 2024. Neural lemmatization and pos-tagging models for coptic, demotic and earlier egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (MLAAL 2024)*, pages 87–97.
- Ollie Sayeed and Bert Vaux. 2017. **66. The evolution of Armenian**, pages 1146–1167. De Gruyter Mouton, Berlin, Boston.
- Chahan Vidal-Gorène, Victoria Khurshudyan, and Anaïd Donabédian-Demopoulos. 2020. **Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing**. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 90–101, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Chahan Vidal-Gorène and Bastien Kindt. 2020. Lemmatization and pos-tagging process by using joint learning approach. experimental results on classical armenian, old georgian, and syriac. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 22–27.

Daniel M. Waldenfels von R. and N. Dobrushina. 2014. Why standard orthography? building the ustya river basin corpus, an online corpus of a russian dialect. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”*, volume 13, pages 720–728.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Appendix

A Detailed Morphological Analysis Results

In MEA, the best-performing features for the RNN model are deixis[psor] and langid with F1-scores of 1.0, tense at 0.968, definite at 0.966, and case at 0.952. However, the model performs poorly on features like polite (0.333), number[psor] (0.333), person[psor] (0.249), degree (0.243), and style (0.124). In MWA, the best features are polarity (0.994), person (0.990), tense (0.988), definite (0.987), and subcat (0.972). The worst tasks in MWA include numform (0.397), number[psor] (0.332), person[psor] (0.249), degree (0.196), and style (0.142). These results highlight the model’s effectiveness in handling certain morphological features while struggling with others, particularly those involving politeness, degree, and stylistic variations. Additionally, Table 4 presents detailed results for the mDeBERTa model.

B Hyperparameters and Experimental Setup

All hyperparameters, the detailed experimental setup and prompts are released in the accompanying GitHub repository to ensure full reproducibility of the experiments.

<i>Feature</i>	<i>CA</i>	<i>MWA</i>	<i>MEA</i>	<i>CA > MEA</i>	<i>CA > MWA</i>	<i>MEA > MWA</i>	<i>MEA > CA</i>	<i>MWA > CA</i>	<i>MWA > MEA</i>
<i>case</i>	0.96	0.98	0.97	0.62	0.64	0.91	0.71	0.70	0.93
<i>number</i>	0.99	0.97	0.97	0.73	0.72	0.89	0.78	0.77	0.93
<i>person</i>	1.00	1.00	0.99	0.95	0.94	0.96	0.89	0.92	0.98
<i>abbr</i>	-	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
<i>adptype</i>	-	1.00	1.00	0.97	0.96	0.99	0.95	0.93	0.99
<i>animacy</i>	1.00	0.97	0.97	0.71	0.71	0.95	0.81	0.75	0.95
<i>aspect</i>	1.00	1.00	0.99	0.91	0.86	0.95	0.88	0.89	0.95
<i>conjtype</i>	-	1.00	-	1.00	1.00	0.99	0.99	1.00	1.00
<i>connegative</i>	-	1.00	1.00	1.00	1.00	1.00	-	1.00	1.00
<i>definite</i>	1.00	0.98	0.98	0.70	0.70	0.95	0.76	0.77	0.96
<i>degree</i>	-	0.99	0.99	0.96	0.96	0.97	0.99	0.99	0.97
<i>deixis</i>	1.00	1.00	1.00	0.98	0.98	0.99	0.90	0.91	0.99
<i>deixis[psor]</i>	-	1.00	-	-	1.00	-	-	-	-
<i>echo</i>	-	-	1.00	1.00	-	-	-	-	1.00
<i>foreign</i>	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.96	0.99
<i>hyph</i>	-	1.00	-	-	1.00	1.00	-	0.99	1.00
<i>langid</i>	-	-	-	-	-	-	-	-	-
<i>mood</i>	0.99	1.00	0.99	0.95	0.89	0.96	0.91	0.90	0.98
<i>nametype</i>	-	0.98	0.99	0.96	0.96	0.98	0.98	0.98	0.99
<i>number[psor]</i>	-	1.00	1.00	0.99	0.99	1.00	0.98	0.98	1.00
<i>numform</i>	-	1.00	1.00	0.97	0.99	1.00	1.00	0.99	0.98
<i>numtype</i>	1.00	1.00	1.00	0.97	0.99	1.00	1.00	0.99	0.99
<i>person[psor]</i>	-	1.00	1.00	0.99	0.99	1.00	0.98	0.98	0.99
<i>polarity</i>	1.00	0.99	0.99	0.86	0.84	0.96	0.87	0.85	0.96
<i>polite</i>	-	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
<i>poss</i>	1.00	1.00	1.00	0.99	0.98	1.00	0.98	0.99	1.00
<i>prontype</i>	0.99	0.99	0.99	0.95	0.92	0.96	0.85	0.87	0.98
<i>reflex</i>	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99	1.00
<i>style</i>	-	0.99	0.98	0.98	0.99	0.99	1.00	0.99	0.98
<i>subcat</i>	0.99	1.00	0.99	0.92	0.90	0.96	0.91	0.91	0.97
<i>tense</i>	-	1.00	-	-	1.00	1.00	-	-	0.99
<i>typo</i>	1.00	1.00	0.99	0.92	0.94	0.95	0.94	0.95	-
<i>verbform</i>	0.99	0.99	0.98	0.84	0.87	0.97	0.87	0.88	0.95

Table 4: Detailed mDeBERTa morphological analysis results for in-domain supervised learning and cross-dialect transfer learning.

Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference

William Thorne[†] Ambrose Robinson[†] Bohua Peng[†] Chenghua Lin[‡]

Diana Maynard[†]

[†] Department of Computer Science, University of Sheffield

[‡] Department of Computer Science, University of Manchester

{wthorne1, bpeng10, d.maynard}@sheffield.ac.uk

ambrose@parablestudio.co.uk

chenghua.lin@manchester.ac.uk

Abstract

As the cultural heritage sector increasingly adopts technologies like Retrieval-Augmented Generation (RAG) to provide more personalised search experiences and enable conversations with collections data, the demand for specialised evaluation datasets has grown. While end-to-end system testing is essential, it's equally important to assess individual components. We target the final, answering task, which is well-suited to Machine Reading Comprehension (MRC). Although existing MRC datasets address general domains, they lack the specificity needed for cultural heritage information. Unfortunately, the manual creation of such datasets is prohibitively expensive for most heritage institutions. This paper presents a cost-effective approach for generating domain-specific MRC datasets with increased difficulty using Reinforcement Learning from Human Feedback (RLHF) from synthetic preference data. Our method leverages the performance of existing question-answering models on a subset of SQuAD to create a difficulty metric, assuming that more challenging questions are answered correctly less frequently. This research contributes: (1) A methodology for increasing question difficulty using PPO and synthetic data; (2) Empirical evidence of the method's effectiveness, including human evaluation; (3) An in-depth error analysis and study of emergent phenomena; and (4) An open-source codebase and set of three llama-2-chat adapters for reproducibility and adaptation.

1 Introduction

The cultural heritage sector is increasingly leveraging advanced technologies like large language models (LLMs) (OpenAI, 2024; Touvron et al., 2023a) and AI assistants (Team Gemini, 2023; Anthropic, 2024) to increase and improve access to collections and their associated data. These technologies provide new opportunities for more dynamic and intuitive interactions with heritage con-

Shark Tank (based on the Dragon's Den reality format) also became a midseason sleeper hit on Sundays in the spring of 2010; the following season, it became the tentpole of the network's Friday night schedule, gradually helping make ABC a strong competitor...

SFT

What reality format did Shark Tank come from?
(answer: the Dragon's Den)

✗ Short Range Dependencies ✗ No Entity Disambiguation ✗ Passage Duplication

PPO

What show became the tentpole of ABC's Friday night schedule? (answer: Shark Tank)

✓ Long Range Dependencies ✓ Entity Disambiguation ✓ Sequence Modification
42 word span > 9 word span "The network" = ABC Shark Tank ↓ show

Figure 1: Example generated questions from supervised-fine-tuned question generation model and one fine-tuned with PPO from synthetic difficulty samples.

tent. One particularly promising technology is Retrieval-Augmented Generation (RAG) (Lewis et al., 2021), which retrieves relevant information from a database of vectorized content to generate accurate, fact-based responses to user queries. We believe that RAG, and iterations on the approach, will play a significant role in improving the search capabilities of heritage institutions in the coming years.

Heritage search systems are used by the public and academics alike; however, the latter tend to submit more complex and specific queries (Koolen and Kamps, 2009). RAG has the capability to fulfil these needs but still requires robust evaluation. This includes not only end-to-end system testing but also the evaluation of individual components. As the response is generally required to be written based *only* on the retrieved documents to mitigate language model hallucinations, we argue that the task is one of Machine Reading Comprehension (MRC). While MRC datasets are well-established in the general domain, they are notably lacking in

cultural heritage and the cost of their construction is prohibitive for most institutions. We estimate that the popular SQuAD dataset cost about \$12,000 to just write the questions, based on their recommended time per question and stated hourly rate of \$9 (Rajpurkar et al., 2016); the actual cost is likely much higher.

To address these challenges, we propose using Automatic Question Generation (AQG) systems to generate MRC datasets. However, we argue that many automatically generated questions, particularly those from zero- or few-shot approaches, do not provide an adequate challenge for modern language models. Manipulating difficulty is challenging through traditional training approaches given its abstract and subjective nature, and prompt based solutions are intractable when considering the infinite permutations and interactions between different aspects of difficulty (Lin et al., 2015a; Rajpurkar et al., 2016; Beinborn et al., 2015; Hsu et al., 2018; Cheng et al., 2021; AlKhuyaey et al., 2023).

To control difficulty, we adapt the Reinforcement Learning from Human Feedback protocol used in AI assistant steering (Ouyang et al., 2022; Bai et al., 2022). In this regime, samples are ranked based on specific criteria and paired into *chosen* and *rejected* samples for training a reward model. This reward model learns to distinguish good samples from bad and outputs a signal which steers a policy model. Rather than relying on costly human annotations, we generate synthetic preference data by evaluating question-answering model performance on a subset of SQuAD, assuming that questions answered correctly less frequently are inherently more difficult. This approach leverages the language model’s innate feature extraction capabilities, eliminating the need to explicitly define difficulty components. Figure 1 demonstrates this feature extraction ability by comparing questions generated with and without reinforcement fine-tuning.

We selected SQuAD over an in-domain QA dataset for two main reasons. First, it is a well-studied, large, and diverse dataset. Second, comparable QA datasets at SQuAD’s scale are either visual question-answering focused (Sheng et al., 2016; Asprino et al., 2022) or have data reliability concerns such as OCR text (Piryani et al., 2024).

This approach enables cultural heritage practitioners to generate challenging evaluation datasets more efficiently and cost-effectively than manual curation. The primary expense is compute resources, which can be accessed in the cloud for

only a few dollars per hour.¹

We summarise this paper’s contributions as follows:

1. A methodology for increasing the difficulty of automatically generated questions using PPO and synthetic data.
2. Empirical evidence of the methodology’s efficacy including human evaluation.
3. An in-depth error analysis and study of interesting phenomena that emerge as part of this approach.
4. An open-source code base and set of models to recreate and adapt our work².

2 Related Work

A similar question generation approach to ours is employed by Zhang et al. (2022) who adopt a pipeline structure. However, their primary objective is to generate suitable questions rather than specifically focusing on difficulty. An important distinction lies in their extensive pre-processing applied to identify candidate answers before feeding them to the question generation model. We argue that pre-identifying answers may limit diversity and prevent the inclusion of potentially complex answer types.

Analyzing and Controlling Question Difficulty. Understanding and managing question difficulty holds significant importance, especially in tasks involving the creation of exams and assessments (Liu and Lin, 2014; AlKhuyaey et al., 2023). One approach, as presented by Loginova et al. (2021), involves modelling the difficulty of multiple-choice questions through the use of softmax scores obtained from a pre-trained QA model. The variance in these scores is then calculated, with higher variance indicating greater difficulty.

Lin et al. (2015b) controls the difficulty of quiz questions through the selection of distractor answers based on semantic similarity between linked data items. This involves collecting both structured RDF data and unstructured text, computing similarity scores through K-means clustering, and generating questions and answers via template-based methods. Importantly, the semantic similarity plays a role in determining the difficulty level, with more

¹<https://huggingface.co/pricing>

²We release all code and a set of three LLaMa-2 adapters on GitHub.

challenging questions featuring distractors exhibiting higher semantic similarity.

Reinforcement Learning with Human Feedback. RLHF is a machine learning paradigm that combines reinforcement learning with human-provided guidance to steer language models to meet the needs of users, finding frequent use in chatbot and AI assistant settings (Ouyang et al., 2022). The basis for most modern methods is the Proximal Policy Optimisation (PPO) algorithm (Schulman et al., 2017), which iteratively enhances the language model’s policy to maximize cumulative rewards through interactions with a dataset or language simulation. It collects experiences, evaluates advantages, and updates the policy with a clipped surrogate objective to ensure stability, gradually improving the model’s performance.

Automatic Question Generation. Chen et al. (2019) introduce a cross-entropy loss with a reinforcement learning-based loss function when training a gated bi-directional neural network for question generation. In this context, the reward model is optimising the semantic and syntactic quality of the question. BLEU-4, as a reward function, optimises the model for the evaluation metrics and the negative Word Movers Distance component is used to ensure semantic quality by maximising the similarity between a generated sequence and a ground truth sequence. Although question quality is maintained, other factors such as question difficulty are not considered.

Self-critic sequence training (SCST) (Rennie et al., 2017) uses a classical policy gradient method, REINFORCE, which is a Monte Carlo method. SCST computes rewards with n-gram token overlap as sub-sentence level rewards. Since training sets often have limited questions, these training rewards are arguably sparse, hindering the question generation model from extrapolating beyond the training distribution. Liu et al. (2019) adopt a two-component reward for refining ill-formed questions. Question wording is used as a measure of short-term reward, and alignment between the question and answer represents a long-term component.

3 Method

To challenge the high cost of manual annotation while maintaining quality and increasing difficulty, we design and implement a robust system capable of generating contextually relevant, coherent, and

challenging question-answer pairs from textual input. The process follows the core methodology of RLHF, deviating only in the use of synthetic preference data to train a reward model. Rather than explicitly defining the characteristics of difficulty and risking failure to capture certain aspects, we exploit the ability of leading question-answer models to derive which questions are challenging, and allow a reward model to extract the component features of the task.

We task three QA models with answering all questions in our validation split of SQuAD. These questions are assigned a score based on the number of times they were answered incorrectly, which are in turn used to generate pairwise preference data. These pairwise samples enable the training of a reward model (RM) for use in fine-tuning a supervised model (SFT) on the task of question generation using Proximal Policy Optimisation (PPO)(Schulman et al., 2017).

We embed this synthetic RLHF process into a greater pipeline for generating samples, shown in Figure 2. This ensures the quality of the final dataset. The pipeline also contains a set of rule-based critics which are used to exclude samples that are malformed and those with non-unique answers in the source text. Samples are then deduplicated using exact string matching.

The remainder of this section discusses each of the relevant components of the pipeline and the RLHF process.

3.1 Supervised Fine-Tuning

In our training process for question generation and response formatting, we begin by employing a reformatted version of the SQuAD v1 training split (see Table 1). The reformatting converts SQuAD to the task of question-answer pair generation, as shown in Figure 3. We select the "correct" answer as the one that appears most frequently in the list of answers for each question in the dataset, selecting randomly among the most common if there is no victor. To ensure model robustness without overfitting, the model undergoes a single epoch of training, enabling it to effectively capture the nuances of the task.

3.2 Reward Modelling

To control the difficulty of our generated questions, we leverage the intrinsic properties present in challenging questions from SQuAD. To extract these attributes, we employ three question answering

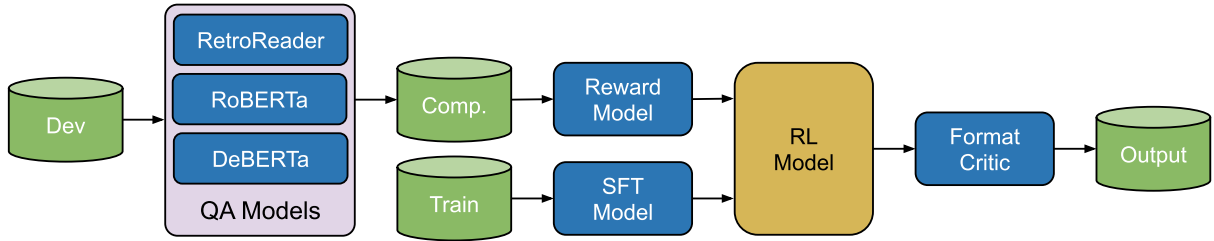


Figure 2: Depiction of our dataset generation pipeline. Question-Answering models are first used to create pairwise comparison data to train a reward model. An SFT model is trained on the train split of SQuAD and then fine-tuned using the reward model, producing the RL model. When generating question-answer pairs for the final dataset, generations are passed through the format critics to ensure data quality.

<p>Instruction Write 1 answerable span extraction question and provide the correct answer based on the text.</p> <p>Input ... Upon its arrival in Canberra, the Olympic flame was presented by Chinese officials to local Aboriginal elder Agnes Shea, of the Ngunnawal people. She, in turn, offered them a message stick ...</p> <p>Response Who received the flame from Chinese officials in Canberra? (answer: <u>Agnes Shea</u>)</p>
--

Figure 3: Example training sample from the reformatted SQuAD dataset for use in supervised fine-tuning.

models that almost match or exceed human performance on SQuAD v2 to evaluate our development split: a RoBERTa-large model³, a DeBERTa-large model⁴ and RetroReader (Zhang et al., 2020). Each question is assigned a score based on the number of models that failed to correctly answer the question. These scores are used to place questions into a pairwise ranking setup against other questions for the same input context. Where a question’s scores are equal, they are considered ties, and no pairwise sample is created. We also record the margin, defined as the difference in score between the chosen and rejected samples, to experiment with the marginal ranking loss, as defined in Touvron et al. (2023b).

3.2.1 Format Critics

To ensure the quality of the final dataset, we utilise a collection of rule-based critics which we call *Format Critics*. These critics have two main functions: they remove questions that don’t adhere to the desired format of $Q?$ (answer: A); they ensure the

provided answer is unique in the text, minimising the number of ambiguous or impossible questions. Samples that pass these critics are then deduplicated using exact matching.

3.3 Reinforcement Training

We use Proximal Policy Optimisation (Schulman et al., 2017) with multiple sets of adapters to reduce the memory overhead during training, implemented using the Transformers Reinforcement Learning library (von Werra et al., 2020). A single base model is used with separate LoRA adapters for the policy, reference, and reward model components; each is switched to perform the relevant aspect of the reinforcement training process.

During early experiments, we found that training was often very unstable or resulted in low pass rates at the format critic. To combat this, we added a rule-based reward component to penalise generations that did not pass the format critic. This simple function converts the reward to be the negative absolute reward in the case that samples are malformed. Using a rule-based reward that manipulates the original reward prevents the instability caused by hard coding a fixed penalty and saves the computational complexity and imperfection of a second adapter-based reward model:

$$R_i = \begin{cases} -|R_i| & \text{if malformed} \\ R_i & \text{otherwise} \end{cases} \quad (1)$$

4 Experimental Setup

4.1 Models

We conduct our experiments with LLaMa2-7B-chat and apply LoRA adapters to all linear layers for all models. This drastically lowers the number of tunable parameters over full-finetuning, enabling training on a single A100 80GB GPU. We also make use of Flash Attention 2 (Dao, 2023) to improve

³deepset/roberta-large-squad2

⁴deepset/deberta-v3-large-squad2

computational and memory efficiency. All LoRA adapters share the same hyperparameters: a LoRA rank of 16, as [Dettmers et al. \(2023\)](#) found rank to have minimal impact on task performance while enabling larger batches through reduced memory usage. This memory efficiency further allowed us to implement sample packing, particularly beneficial with Flash Attention 2’s preference for minimal padding. We set alpha to twice the rank ⁵, use a dropout of 0.05 - shown optimal for 7B models by [Dettmers et al. \(2023\)](#), and maintain LLaMa-2’s BF16. As a baseline, we compare to LLaMa-2-7B-chat in a zero-shot setting (see Appendix B).

We experiment with marginal ranking loss to help distinguish between slight and significant differences in question difficulty while training the reward model. Under the hypothesis that the difficulty of a question is not independent of the associated passage of text, we also experiment with training a reward model with and without the input text attached. Results of these experiments can be found in Appendix A.

4.2 Generation Settings

During generation, the model is tasked with producing a single output for each question in the training set using nucleus sampling ([Holtzman et al., 2020](#)). We maintain the original configuration for LLaMa-2 with a repetition penalty of 1.1, top P of 0.7, and top K of 0 but increase the temperature from 0.6 to 0.9 to increase the diversity of generations.

4.3 Data Splits

We base our splits off the original SQuAD to minimise the risk of data leakage. We maintain the full train split unchanged as any model previously trained on SQuAD will have seen the full train split. We extract a test split of 500 contexts from the dev split, ensuring no contexts appear in both the dev and test splits. We extract 50 unique contexts from the test split for a human evaluation of question quality and answerability. In all cases, context-question pairs were only kept if they fit into the context length of LLaMa-2 when formatted in the correct prompt format. All samples were formatted into the three instruction components: *instruction*, *input*, *response* as shown in Figure 3.

Only the dev set of our SQuAD dataset was used to derive difficulty comparison data, to ensure the reward model never sees the samples used for eval-

Split	# Contexts	# Questions
Train	18,891	87,599
Dev	1,567	8,038
Test	500	2,532
Human Test	50	50
Train comp.	1,107	8,394
Dev comp.	123	950

Table 1: Split of contexts and questions from SQuAD. The *comp.* splits are derived from the dev split, used to evaluate the performance of the reward model during training.

uation. To evaluate the reward model, we extract 10% of the comparison contexts. Full dataset statistics can be found in Table 1.

4.4 Evaluation Metrics

As our goal is to evaluate the difficulty of answerable questions, we provide the input passage, question and answer to GPT-4o⁶ and Gemini-1.5-pro⁷ and ask whether the sample meets our specification of validity. We take samples to be answerable if they were unanimously labelled as such, and reject all other samples. GPT-based evaluations have demonstrated a robust alignment with human preferences across various complex tasks in reference-free settings ([Fu et al., 2023](#); [Liu et al., 2023](#)). The results of this analysis can be found in Appendix C.

To assess the quality of generated questions relative to our SQuAD test split, we *intentionally avoid* *n*-gram based metrics such as BLEU ([Papineni et al., 2002](#)), ROUGE ([Lin, 2004](#)), and more modern alternatives such as Q-Metrics ([Nema and Khapra, 2018](#)), as we believe they restrict diversity of generation, constraining the model to reference questions and answers. We instead adopt the following reference-free metrics:

Syntactic Divergence provides a distance measure between two dependency paths which acts as a measure of difficulty. Word-lemma anchors, common to both the question and answer sentence, are first detected. A dependency path from the anchor to the interrogative word (who, what, etc.) in the question is compared to the dependency path between the anchor and the answer span in the answer sentence using Levenshtein distance ([Levenshtein et al., 1966](#)).

RQUGE calculates an *acceptability-score* by generating an answer for the candidate question and predicting the semantic similarity between the

⁵<https://lightning.ai/pages/community/lora-insights/#toc9>

⁶gpt-4o as of 1st June 2024

⁷gemini-1.5-pro as of 1st June 2024

Model	Total Valid (\uparrow)	DeBERTa (\downarrow)	RoBERTa (\downarrow)	RetroReader (\downarrow)
SQuAD	2,532 (-)	0.68	0.68	0.65
ZeroShot	357 \pm 14 (0.14)	0.644 \pm 0.007	0.650 \pm 0.007	0.629 \pm 0.009
SFT	1252 \pm 2 (0.49)	0.654 \pm 0.012	0.653 \pm 0.005	0.616 \pm 0.015
PPO-input	1375 \pm 18 (0.54)	0.601 \pm 0.004	0.606 \pm 0.003	0.582 \pm 0.007
PPO-input-margin	1373 \pm 4 (0.54)	0.612 \pm 0.001	0.608 \pm 0.005	0.587 \pm 0.002

Table 2: Question-Answering model performance on each set of samples. Models were only supplied samples which passed the format critics and were unanimously deemed answerable by GPT-4o and Gemini-1.5-pro. The *Total Valid* column indicates this number of valid samples used during question answering. Accuracy is based on exact match and results are mean and standard deviation across three sets of generated samples. Lower accuracy indicates harder questions.

predicted answer and the gold answer provided by the user. In our setup, this metric acts as an assessment of both the question and answer quality (Mohammadshahi et al., 2023).

QAScore attempts to align AQG evaluation to human judgements. Question-answer pairs are evaluated by summing log-probabilities of RoBERTa correct token predictions for all words in the answer when masked individually. QAScore claims to show strong correlation with human judgement (Spearman $r = 0.864$) (Ji et al., 2022).

Self-BLEU assesses how similar questions are to other questions generated for a given context. Each question is taken as a hypothesis and the others as a reference for the BLEU calculation. The self-BLEU is taken as the average BLEU for the question collection (Zhu et al., 2018).

5 Results and Discussion

Model Accuracy. To measure performance, we observe the difference in prediction accuracy for QA models on each dataset. Table 2 shows that in all cases of PPO training, we observe a decrease in average model prediction accuracy and an increase in the total number of valid generations. The consistent decrease in absolute prediction accuracy for all models when using the PPO trained models over both zero-shot and SFT signifies an increase in average question difficulty. The SFT process vastly improves the model’s ability to generate valid questions. The PPO process further bolsters this capability which illustrates that the model is learning the intrinsic properties of high-quality questions. The performance of the reward models, shown in Appendix A, is reflected here, showing lesser degrees of improvement for those models fine-tuned without access to the input passage.

External Metrics. Figure 4 shows results for the reference-free metrics. RQUGE is clearly effective

at discriminating between human-written SQuAD samples, those generated by the fine-tuned models and the zero-shot examples, but it is unable to separate out the SFT and PPO results. The particularly high score for SQuAD could in part be due to data leakage as the answer generation model for the metric was trained on SQuAD (Khashabi et al., 2022). This would indicate why our newly generated questions might score lower as it cannot have memorised the answer. Syntactic divergence results for the SQuAD test split and all trained model generations follow a consistent distribution but the zero-shot results appear much better, despite having a higher average prediction accuracy than the SFT and PPO models. Zero-shot obtaining higher syntactic divergence could stem from the general purpose language generation objective of LLaMa-2-chat. This could cause the model to generate boilerplate text which distances the structure of the question from that of the answer sentence but doesn’t necessarily result in a more difficult question. QAScore proves uninformative, only being able to subtly identify SQuAD samples from model generated samples. Self-BLEU indicates that SQuAD samples are the most diverse, which is to be expected, but that zero-shot samples exhibit a distinct lack of diversity when compared with fine-tuned models. This result is, in part, misleading as Self-BLEU was only calculable for input passages with at least two valid questions. As the number of valid generations was so low for the zero-shot model, the cases where multiple valid questions were generated for a context was disproportionately in favour of identical generations.

In general we find the reference-free metrics to show limited correlation with model prediction accuracy and an ability to differentiate human written samples from model generations. We believe this is evidence for the continued need for more reli-

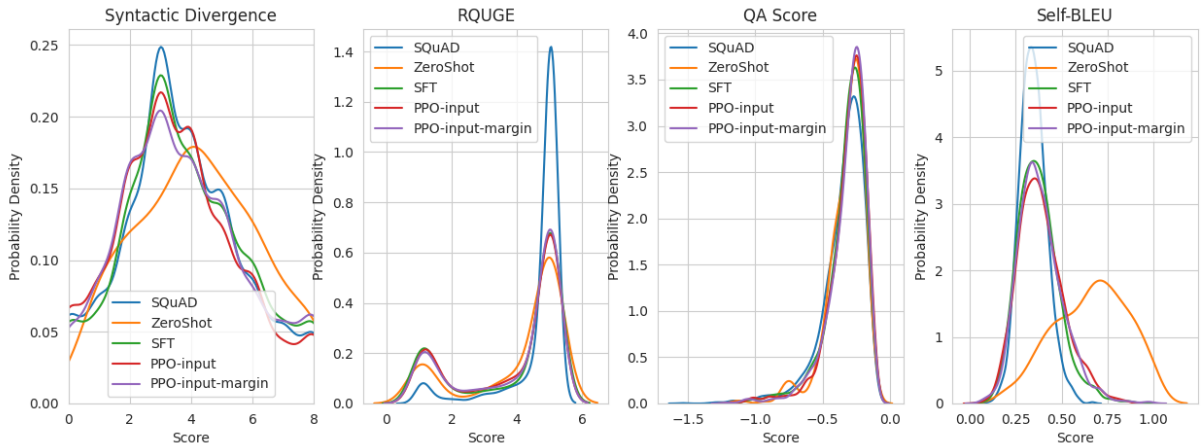


Figure 4: Distribution of reference free metrics results for each model’s generations based on our SQuAD test set.

able, reference-free evaluation tools for question generation.

Human Evaluation. To evaluate question quality, we conduct a human evaluation on a subset of 50 passages from the test split. Each input passage and question is filtered through the format critic then provided to two annotators who select either the correct answer span or indicate that the question cannot be answered. In the case of annotator disagreement or the annotated answers differing from the model generated answer, the annotator responses and the model answer are provided to two new annotators who both select which responses are appropriate. We allow annotators to select multiple responses as correct but only include those that were selected unanimously by both annotators as valid. We observe an agreement of $\kappa = 0.7975$ between annotators. The results of this evaluation, shown in Table 3, displays an equivalent or improved rate of answerability when fine-tuning with PPO; the answerability proportions for each dataset are roughly equivalent to those presented in Table 2. This further corroborates the efficacy of our approach.

The results demonstrate that reinforcement learning can effectively manipulate question difficulty, while highlighting important avenues for future work. While SQuAD’s synchronic nature served our experimental needs, cultural heritage datasets typically present diachronic challenges that add complexity to question generation. Although specialised diachronic models exist [Drinkall et al. \(2024\)](#), they lack the extensive training of general-domain LLMs. However, these larger models’ exposure to historical corpora, combined with their advanced instruction-following capabilities, sug-

Model	Full	Partial
ZeroShot	0.10	0.14
SFT	0.52	0.60
PPO-input	0.52	0.64
PPO-input-margin	0.56	0.64

Table 3: Results of human evaluation for question quality. *Full* indicates that the model generated answer was a valid answer according to the format critics and identified by human annotators and *Partial* indicates that the sample passed format critics and a valid answer was identified for the question but the model generated answer did not match.

gests potential for manipulating temporal complexity as an additional dimension of question difficulty.

5.1 Error Analysis

Failure Modes. At a high level, we can observe the reasons for sample rejection for each model. As shown in Figure 5, the zero-shot model is generally unable to generate samples that have a single answer span in the text, despite exactly specifying this in the prompt. The high number of incorrectly formatted samples was a result of only a question being generated or neither a question nor answer being generated. For all the trained model variants, the dominant failure mode was unanswerable questions. As shown in Appendix C, each of the fine-tuned models show a similar proportion of otherwise valid samples being unanswerable. The answerability rate could potentially be improved by generating candidate answers, as in ([Zhang et al., 2022](#)), and passing an input passage and answer to the question generation model.

Positional Bias. One interesting phenomenon is the positional bias in where the model chooses to

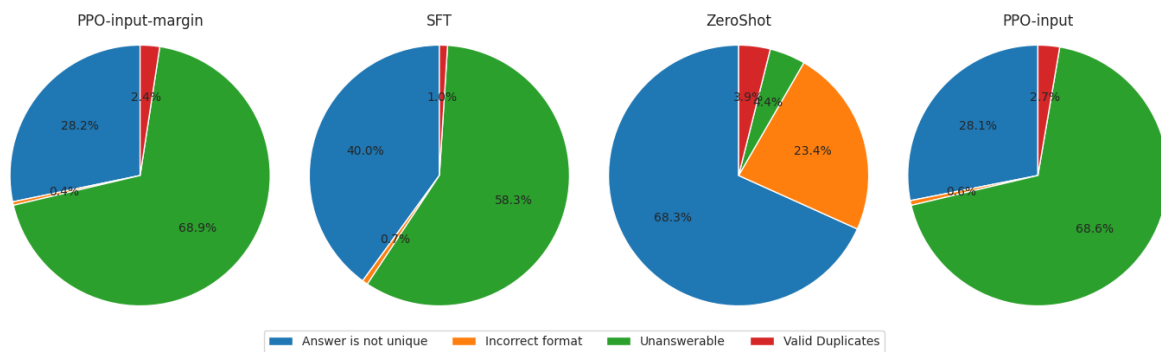


Figure 5: Error distribution of questions for SFT, ZeroShot, and the two best performing PPO variants.

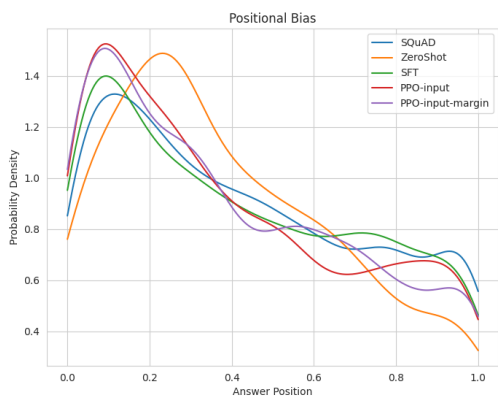


Figure 6: Position of answer span, merged to be a single word, as a proportion of the way through the input passage when split into words. SQuAD positions are selected from our test split and answers are chosen to be the most common from the list of suitable answers. Neither invalid nor exact duplicate questions are considered.

generate answers. To calculate positional bias, we treat the full answer span as a single "word" and calculate the proportion through the input paragraph in which the answer word appears. As seen in Figure 6, the zero-shot positional bias is less severe than in the other datasets. The positional bias of SQuAD is clearly seen as, after training on the dataset, all models exhibit this same preference for the beginning of input passages. The clear bias observed in the zero-shot model, despite not being fine-tuned, is documented in other tasks such as LLM ranking (Wang et al., 2023; Li et al., 2023) and in summarisation where introductory content is favoured (Ravaut et al., 2023). A potential remedy is to supply the model with a sliding window of sentences across the context paragraph to force the model to generate questions throughout the text.

While this would improve the diversity of a final dataset, it may have the adverse effect of limiting the range of dependencies, restricting potentially challenging questions across the whole text.

Hallucinated External Knowledge. Where ambiguous references to specific entities exist in the input passage such as *the museum collection*, the models frequently attempt to fill in which entity is being referred to. From a context containing ambiguous references to an unnamed museum, the questions *What year did the Tate acquire the statue of St John the Baptist?*, *How many works does Rodin have in the British Museum's collection?* were generated across both the SFT and PPO models; the examples consistently passed LLM evaluations of answerability. This suggests the solution to this problem is more holistic and requires improvements at a foundational model level to resolve. We could resolve this at a critic level through more careful prompting, however, this returns to our original and intractable task of textually describing a complex task. A more holistic solution could be to adapt PPO with functional grounding (Carta et al., 2023) to be a pure text task. However, this may lower the quality of questions as it could discourage the use of implicit or complementary knowledge.

Unidirectional Relationships. A strategy to increase the difficulty of questions is to invert relationships found in the text. The models sometimes misappropriate this tool, resulting in invalid questions such as the question *What did the Ming dynasty represent?* from a passage containing *...explorer Zheng He representing the Ming Dynasty...* Knowledge graph assisted generation could help to resolve these logical inconsistencies (Lin et al., 2015a). However, expecting our target demograph-

ics, emerging domains, to possess high-quality knowledge graphs is an unreasonable assumption.

6 Conclusion

In this paper, we introduce a low-cost methodology for generating challenging MRC datasets to meet the growing need for evaluation datasets in the cultural heritage sector. By using high-performing question-answering models to identify the most difficult questions, we were able to create synthetic pairwise data for training a reward model. Rather than manually defining question difficulty, our approach allows the model to learn and extract these features autonomously, leading to a significant improvement in the difficulty of questions generated for evaluation.

With this said, we trained on a general domain dataset in order to single out the training behaviour, in doing so losing many of the characteristic features of heritage datasets. In future work we will examine how the training paradigm fares under the unique challenges presented by such a varied industry.

Although this work was produced to meet the evaluation demands of our ongoing work in RAG at our institution, we also highlight that the approach can work in any domain and that with some modification, it could be used to augment other dataset formats. We believe this approach can be extended further, allowing for the manipulation of multiple abstract properties simultaneously through multi-reward model setups (Wu et al., 2023).

Limitations

This project only shows the suitability of the method on a single model. In future work, we seek to address this by performing a more comprehensive review of the approach across a range of model sizes and architectures. We also acknowledge that this method currently only addresses answerable questions while most contemporary QA datasets utilise both answerable and unanswerable questions. Finally, despite using LoRA and multi-adapter training, we still required approximately 15 GPU hours on an A100 80GB which restricts the potential audience for this approach. Evaluating smaller models or quantisation will enable greater access to this project's benefits.

Acknowledgements

This work was supported by the Arts and Humanities Research Council [grant number AH/X004201/1].

Ethics Statement

This project has been approved by the relevant institution's ethics committee. We use LLaMa2 in accordance with Meta's license⁸. All annotators were located through word of mouth and paid £12 per hour - above the UK National Living Wage of £11.44.

References

- Samah AlKhuzayy, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2023. [Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches](#). *International Journal of Artificial Intelligence in Education*.
- Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.
- Luigi Asprino, Luana Bulla, Ludovica Marinucci, Misaël Mongiovi, and Valentina Presutti. 2022. [A Large Visual Question Answering Dataset for Cultural Heritage](#). In *Machine Learning, Optimization, and Data Science*, pages 193–197, Cham. Springer International Publishing.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. [Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback](#).
- Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. [Candidate evaluation strategies for improved difficulty prediction of language tests](#). In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.
- Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. [Grounding large language models in interactive environments with online reinforcement learning](#).

⁸<https://ai.meta.com/llama/license/>

- Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2019. [Reinforcement learning based graph-to-sequence model for natural question generation](#). *CoRR*, abs/1908.04942.
- Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. [Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978.
- Tri Dao. 2023. [FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning](#). ArXiv:2307.08691 [cs].
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient Finetuning of Quantized LLMs](#).
- Felix Drinkall, Eghbal Rahimikia, Janet B. Pierrehumbert, and Stefan Zohren. 2024. [Time Machine GPT](#).
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. [GPTScore: Evaluate as You Desire](#). ArXiv:2302.04166 [cs].
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The Curious Case of Neural Text Degeneration](#). ArXiv:1904.09751 [cs].
- Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. [Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques](#). *Information Processing & Management*, 54(6):969–984.
- Tianbo Ji, Chenyang Lyu, Gareth Jones, Liting Zhou, and Yvette Graham. 2022. [QAScore—An Unsupervised Unreferenced Metric for the Question Generation Evaluation](#). *Entropy*, 24(11):1514.
- Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. [UnifiedQA-v2: Stronger Generalization via Broader Cross-Format Training](#). ArXiv:2202.12359 [cs].
- Marijn Koolen and Jaap Kamps. 2009. [Information Retrieval in Cultural Heritage](#). *INTERDISCIPLINARY SCIENCE REVIEWS*, 343:268–284.
- Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#).
- Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023. [Split and Merge: Aligning Position Biases in Large Language Model based Evaluators](#). ArXiv:2310.01432 [cs].
- Chenghua Lin, Dong Liu, Wei Pang, and Edward Apeh. 2015a. [Automatically Predicting Quiz Difficulty Level Using Similarity Measures](#). In *Proceedings of the 8th International Conference on Knowledge Capture, K-CAP 2015*, pages 1–8.
- Chenghua Lin, Dong Liu, Wei Pang, and Zhe Wang. 2015b. [Sherlock: A semi-automatic framework for quiz generation using a hybrid semantic similarity measure](#). *Cognitive computation*, 7:667–679.
- Chin-Yew Lin. 2004. [ROUGE: A Package for Automatic Evaluation of Summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Dong Liu and Chenghua Lin. 2014. [Sherlock: a semi-automatic quiz generation system using linked data](#). In *ISWC (Posters & Demos)*, pages 9–12.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment](#). ArXiv:2303.16634 [cs].
- Ye Liu, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2019. [Generative question refinement with deep reinforcement learning in retrieval-based QA system](#). *CoRR*, abs/1908.05604.
- Ekaterina Loginova, Luca Benedetto, Dries Benoit, and Paolo Cremonesi. 2021. [Towards the Application of Calibrated Transformers to the Unsupervised Estimation of Question Difficulty from Text](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 846–855, Held Online. INCOMA Ltd.
- Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. [RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada. Association for Computational Linguistics.
- Preksha Nema and Mitesh M. Khapra. 2018. [Towards a Better Metric for Evaluating Question Generation Systems](#).
- OpenAI. 2024. [GPT-4 Technical Report](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. [Training language models to follow instructions with human feedback](#). *Advances in Neural Information Processing Systems*, 35:27730–27744.

- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. [Bleu: a Method for Automatic Evaluation of Machine Translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. [ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages](#).
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ Questions for Machine Comprehension of Text](#). ArXiv:1606.05250 [cs].
- Mathieu Ravaut, Shafiq Joty, Aixin Sun, and Nancy F. Chen. 2023. [On Context Utilization in Summarization with Large Language Models](#). ArXiv:2310.10570 [cs].
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal Policy Optimization Algorithms](#). ArXiv:1707.06347 [cs].
- Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. 2016. [A Dataset for Multimodal Question Answering in the Cultural Heritage Domain](#). In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17, Osaka, Japan. The COLING 2016 Organizing Committee.
- Team Gemini. 2023. [Gemini: A Family of Highly Capable Multimodal Models](#). <https://arxiv.org/abs/2312.11805v4>.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). ArXiv:2307.09288 [cs].
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. [Trl: Transformer reinforcement learning](#). <https://github.com/huggingface/trl>.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. [Large Language Models are not Fair Evaluators](#). ArXiv:2305.17926 [cs].
- Zequ Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. [Fine-Grained Human Feedback Gives Better Rewards for Language Model Training](#). ArXiv:2306.01693 [cs].
- Cheng Zhang, Hao Zhang, Yicheng Sun, and Jie Wang. 2022. [Downstream transformer generation of question-answer pairs with preprocessing and post-processing pipelines](#). In *Proceedings of the 22nd ACM Symposium on Document Engineering, DocEng '22*, pages 1–8, New York, NY, USA. Association for Computing Machinery.
- Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020. [Retrospective reader for machine reading comprehension](#). *CoRR*, abs/2001.09694.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Texygen: A Benchmarking Platform for Text Generation Models](#). ArXiv:1802.01886 [cs].

A Reward Model Performance

To understand the relative contributions of marginal ranking loss and the use of the input when training

Model	Accuracy (%)
RM	63.66
RM-input	70.69
RM-margin	62.39
RM-input-margin	70.38

Table 4: Accuracy of reward model variants based on the test split of the comparisons dataset. *input* indicates that the model was trained with the question and associated text passage as input and *margin* indicates that marginal ranking loss was used.

reward models to discriminate based on difficulty, we trained all four permutations of settings on the whole training split of the comparisons dataset and evaluated on the test split. As shown in Table 4, the inclusion of the input text had a very significant impact on performance. This was expected as the difficulty of a question is not independent of the related passage. Surprisingly, marginal ranking loss had a very slight negative impact on reward model performance. We believe this could be due to the fact that features of difficulty are very subtle and the marginal component may have caused too significant adjustments due to higher loss values.

B Obtaining Zero-Shot Model Generations

To obtain zero-shot generations, we adopt a slightly different approach. To avoid overconstraining the output of the model, we adopted a two-stage process. LLaMa-2-7b-chat was first tasked with generating a question-answer pair based on the text, unconstrained. We then passed this output back into the model with the task of extracting the question and answer components and placed them into a JSON file with the keys *question* and *answer*. We used the same, high temperature of 0.9 for generating the samples and a much lower temperature of 0.2 for extracting into a JSON to reduce the chance of models altering the generated sequences while structuring them.

C API-Based LLM Answerability Annotation

To ensure that we evaluate performance on as high-quality questions as possible, we extract only those questions deemed *answerable*, by our definition, by both GPT-4o and Gemini-1.5-pro. Table 5 shows that the zero-shot samples had the highest rate of predicted answerability; each other variant shows

Following is a text, a question and an answer. You must determine whether the provided answer is a correct span-extraction response to the question. If there are multiple plausible answers in the text, the answer should be the most relevant or accurate one. If there are multiple equally plausible answers in the text, respond "NO". If the provided answer is incomplete or contains excess information, respond "NO". If the answer does not correctly answer the question, respond "NO". Only if the answer is correct and does not breach the aforementioned requirements, respond with "YES".

Text: ... Upon its arrival in Canberra, the Olympic flame was presented by Chinese officials to local Aboriginal elder Agnes Shea, of the Ngunnawal people. She, in turn, offered them a message stick ...

Question: Who received the flame from Chinese officials in Canberra?

Answer: Agnes Shea

Respond with only "YES" or "NO" in response to this task. Do NOT provide any other text or reasoning.

Figure 7: Example prompt and response to GPT-4o (gpt-4o as of 1st June 2024) and Gemini-1.5-pro (gemini-1.5-pro as of 1st June 2024).

Model	Answerable (\uparrow)	Unanswerable (\downarrow)	Undetermined (\downarrow)	Cohen's κ (\uparrow)
ZeroShot	0.73	0.14	0.13	0.61
SFT	0.64	0.20	0.16	0.62
PPO	0.64	0.20	0.16	0.62
PPO-input	0.62	0.20	0.18	0.58
PPO-margin	0.62	0.19	0.19	0.56
PPO-input-margin	0.63	0.21	0.16	0.62

Table 5: Results of answerability task posed to GPT-4o and Gemini-1.5-pro. Results represent the proportion of samples that are answerable, unanswerable and undecided, taken from those samples which passed the format critic.

very consistent rates of answerability. This outcome should be tempered by the results in Figure 5 which indicates that the zero-shot model had an extremely high failure rate in many other regards.

Assessing Large Language Models in Translating Coptic and Ancient Greek Ostraca

Audric-Charles Wannaz

University of Basel
Petersgraben 51
4051 Basel, Switzerland
audric.wannaz@unibas.ch

So Miyagawa

University of Tsukuba
1-1-1 Tennodai
Tsukuba, Ibaraki, Japan
miyagawa.so.kb@u.tsukuba.ac.jp

Abstract

The advent of Large Language Models (LLMs) substantially raised the quality and lowered the cost of Machine Translation (MT). Can scholars working with ancient languages draw benefits from this new technology? More specifically, can current MT facilitate multilingual digital papyrology? To answer this question, we evaluate 9 LLMs in the task of MT with 4 Coptic and 4 Ancient Greek ostraca into English using 6 NLP metrics. We argue that some models have already reached a performance that is apt to assist human experts. As can be expected from the difference in training corpus size, all models seem to perform better with Ancient Greek than with Coptic, where hallucinations are markedly more common. In the Coptic texts, the specialised Coptic Translator (CT) competes closely with Claude 3 Opus for the rank of most promising tool, while Claude 3 Opus and GPT-4o compete for the same position in the Ancient Greek texts. We argue that MT now substantially increases the incentive to work on multilingual corpora. This could have a positive and long-lasting effect on Classics and Egyptology and help reduce the historical bias in translation availability. In closing, we reflect upon the need to meet AI-generated translations with an adequate critical stance.

1 Introduction

Translations have been the cornerstone of scholarly activity in the fields of Classics and Egyptology since their inception, serving both academic and public dissemination purposes (Balmer, 2009; Westerfeld, 2016). The initial preference for Latin as the target language for translations reflects its status as the scholarly *lingua franca* during the early phase of these disciplines (Lockwood, 1918; Burke, 2017). Over the centuries, there has been a steady transition to vernacular languages in order to make scientific content more accessible for an attempt to partly reverse this transition; See Merisalo

2015 for the example of the Italian language). In contemporary practice, English has emerged as the preferred *lingua franca*, broadening the accessibility and scope of translated texts (Nørgaard, 1958). This article focuses specifically on English translations in a field tangent to both Classics and Egyptology, namely digital documentary Papyrology¹. In this specific area of study, interdisciplinary communication between Classics and Egyptology have improved slightly in last decades (van Minnen, 1993, 14). However, English translations of primary sources are not yet widely available on the Web, as the next section shows.

Currently, virtually all available translations of ancient Greek and Coptic texts have been made by human experts. The coverage of those translations reveals major disparities. On Papyri.info² as of 5 May 2024, 59,955 Greek texts with transliteration are available, but only 5,678 are accompanied by translations in English, and 628 in other languages, i.e. around one tenth of the total corpus. The situation is even more critical for Coptic, where out of 2,099 texts, 2,049 are untranslated, and only 50 texts are available in English, French or German, that is less than one per cent of the total corpus. If other translations exist, they are mainly printed and are not easily accessible online, making them unsuitable for research in digital papyrology.

Meanwhile, the landscape of AI-generated translations has evolved considerably, from simple rule-

¹“Digital Papyrology can be defined as the whole set of electronic resources and methodologies aimed at creating, storing, accessing, processing, and publishing information pertaining to research and study in the various fields of interest of the papyrological discipline.” (Reggiani, 2017, 8).

²“Papyri.info has two primary components. The Papyrological Navigator (PN) supports searching, browsing, and aggregation of ancient papyrological documents and related materials; the Papyrological Editor (PE) enables multiauthor, version controlled, peer reviewed scholarly curation of papyrological texts, translations, commentary, scholarly metadata, institutional catalog records, bibliography, and images.” <https://papyri.info/> [Accessed: 24/05/2024].

based systems to sophisticated machine learning models. Early efforts in computer-aided translation were fundamental, but limited in terms of accuracy and scope. In recent years, the adoption of machine learning models has significantly improved the quality of translations. Since its launch in 2006, Google Translate (Wu et al., 2016) has long been the benchmark for machine translation tools, despite the initial lack of support for languages such as Coptic or ancient Greek.³ More recently, DeepL, introduced in 2017, has set new standards for the accuracy of machine translations (although it does not include the languages in question). The most advanced development in this area concerns large generative multimodal language models (LLMs), which are serious contenders for complex translation tasks (Yang et al., 2024; Gaspari, 2024).

1.1 Research Question

1.1.1 General Research Problem

Given the obvious gaps in translation in the fields of Classics and Egyptology, especially with respect to Digital Papyrology, this study will investigate whether MT can effectively fill these gaps today. The central question concerns the ability of modern AI-driven tools to provide accurate and reliable translations for ancient documents that remain largely untranslated or not digitised.

1.1.2 Specific Research Objectives

The objective of this study is to evaluate the effectiveness of MT systems in facilitating multilingual digital papyrology. This includes a comprehensive examination of the performance of these technologies in translating Coptic and Ancient Greek, two common languages in this field (Vierros and Henriksson, 2017; Dahlgren, 2018). To this end, our methodology is structured as follows:

Evaluation of the MT of Coptic texts (Section 2): We first introduce four Coptic texts (2.1). Then, we introduce 6 NLP metrics to evaluate the performance of 9 LLMs: the Coptic Translator, a LLM specialised on the task of Coptic-English as well as 8 generic LLMs (2.2). Evaluation of Ancient Greek texts (Section 3): Similarly, four Ancient Greek texts are presented as close equivalents to the four Coptic texts in form and content (3.1). In a second step, we describe how the same LLMs used to translate Coptic texts fare in the same task in this

³Generally, the production of textual corpora, which can be training data for machine translation, has been much less in Coptic than in Greek; cf. Clackson (2004).

other ancient language (3.2). Comparative analysis (Section 4): on the basis of all produced results of MT for Coptic and Ancient Greek, we discuss the overall impact of AI-generated translations on the field of digital papyrology. Future directions (Section 5): The study closes with a discussion of the potential future implications of integrating MT into academic research and public dissemination.

In sum, the aim of this pilot study is to provide some empirical information on the current practical capabilities of AI in translating ancient texts and to stimulate debate on its strategic integration in the fields of Classics and Egyptology.

2 Evaluating Coptic-English MT

2.1 Four Coptic Texts

To evaluate the performance of Machine Translation (MT) on Coptic texts, we selected four relatively well-preserved documentary letters written on ostraca from the IFAO (Institut français d'archéologie orientale) collection: TM 874362, 874363, 874364, and 874365, which are unlikely to be used in the training of the existing LLMs.⁴

- TM 874362/ IFAO OC 252 (C 1906): 11 x 11 cm. VII CE, Western Thebes. Late Roman Amphora 7, Letter from Petros concerning a church vessel of Apa Menas in Ape (Luxor).
- TM 874363/ IFAO Inv. OC 275 (C 1917): 16 x 10 x 1.2 cm. VII-VIII CE, Theban region. Late Roman Amphora 7, Letter from the sick Antonios to Petros, asking for money, possibly to buy medicine.
- TM 874364/ IFAO Inv. OC 104 (C 1916): 10 x 9.5 cm. VII CE, Theban region. Letter from Psmoei to a deacon announcing the repayment of a tremissis and requesting lentils. Fragment of a red Pseudo-Aswanese Late Roman Amphora.
- TM 874365/ IFAO Inv. OC 270 (C 1879): 17 x 11 cm. VII-VIII CE, Thebes(?). Pseudo-Aswanese pottery. Letter concerning exchange of crops, vegetables, dates, arax (legume), and oil between several individuals.

⁴Later, more comprehensive studies will benefit from using a larger sample size. This preliminary study chooses to limit its scope to a few short texts that are rather homogeneous in content.

These texts were chosen to cover a range of preservation states (TM 874362 and TM 874365 are well-preserved, TM 874364 is sufficiently well-preserved, TM 874363 is partially preserved) and standardized character lengths (averaging about 225 characters). Ground truth reference translations were produced by Coptic scholars under the supervision of two eminent experts (Anne Boud’hors and Esther Garel; see [Boud’hors and Garel, 2019](#)).

2.2 Assessing Coptic-English MT

2.2.1 LLMs

We compared the output of 9 LLMs on the MT task: The dedicated Coptic Translator model. GPT model family:⁵ GPT-4o, GPT-4, GPT-3.5. Claude model family: Claude Opus, Claude Sonnet, Claude Haiku. Gemini model family: Gemini Advanced, Gemini.⁶

The Coptic Translator ([Enis and Megalaa, 2024](#)), developed by Maxim Enis and Andrew Megalaa from Williams College Computer Science Department, is the first contextual machine translation system for the Coptic language. The authors created the system by fine-tuning pretrained multilingual transformer models on limited Coptic-English parallel data and employing techniques such as romanization, back-translation, and transfer learning, resulting in strong translation performance on religious Coptic texts. The translator provided the first-ever English translations for over 5,700 previously untranslated Coptic sentences and will be open-sourced and made freely available online to assist Coptic language learners, scholars, and those working to revive the language.

2.2.2 Metrics

To quantitatively assess translation quality, we employed 6 metrics from the field of natural language processing (NLP):

- “school”: a custom metric designed to mimic a human approach to the task of MT evaluation (see [Figure 1](#)).⁷

⁵Models within a family are listed in decreasing recency. Model size and performance are generally correlated.

⁶Since the conception of this paper, several significant LLMs have been released, including OpenAI’s o1-mini and o1-preview. These newer models will be incorporated into future studies, with the goal of generating more robust quantitative results compared to the preliminary findings presented in this proof-of-concept study.

⁷Similar to a teacher correcting a test at school, this metric counts “mistakes” (words absent either from the base or tar-

- Levenshtein distance: character-level edit distance.
- BLEU (BiLingual Evaluation Understudy; [Papineni et al. 2002](#)): n-gram precision with a brevity penalty. To add reliability, we used the standardised version, SacreBLEU ([Post, 2018](#)).
- TER (Translation Error Rate; [Snover et al., 2006](#)): word-level edit distance.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering; [Banerjee and Lavie, 2005](#)): alignment-based metric.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation; [Lin, 2004](#)): n-gram recall.

```
import difflib

def school_metric(base_text, text,
    ↪ weights={'reused_diff': 0.5, '
    ↪ not_reused_or_present': 1}):
    words1, words2 = set(base_text.split
    ↪ ()), set(text.split())
    #SequenceMatcher algorithm
    reused_diff = sum(1 for w1 in words1
    ↪ if difflib.get_close_matches
    ↪ (w1, words2, n=1, cutoff=0.8)
    ↪ and w1 not in words2)
    not_reused_or_present = len(words1.
    ↪ symmetric_difference(words2))

    score = (weights['reused_diff'] *
    ↪ reused_diff +
    ↪ weights['
    ↪ not_reused_or_present
    ↪ '] *
    ↪ not_reused_or_present
    ↪ )

    return score
```

Figure 1: Python code of “school” scoring

These metrics capture different aspects of similarity between the MT and human reference translations. Levenshtein and TER measure the amount of editing required to transform one text into the other. BLEU, METEOR (optimised to evaluate MT) and

get text) and “half-mistakes” (words reused in target text but slightly modified).

metric	4 Coptic Texts		4 Greek Texts	
	top3models	meanscore	top3models	meanscore
school	gemi (71.75)	58.14	gpt_3.5 (45.38)	37.03
	gemi_advanced (67.00)		gemi (40.88)	
	gpt_4o (61.25)		gpt_4 (38.50)	
levenshtein	gemi_advanced (312.50)	223.78	gpt_3.5 (188.25)	148.16
	gpt_4 (237.25)		gpt_4 (157.25)	
	gemi (236.75)		gemi (156.75)	
ter	gemi_advanced (1.27)	0.92	gpt_3.5 (0.63)	0.50
	gpt_4 (0.99)		gemi (0.55)	
	gemi (0.95)		gpt_4 (0.53)	
sacrebleu	claude_opus (20.02)	5.98	gpt_4o (39.63)	30.89
	claude_haiku (11.52)		claude_opus (37.18)	
	coptic_translator (8.43)		claude_sonnet (33.89)	
meteor	claude_opus (0.46)	0.23	claude_opus (0.67)	0.60
	claude_haiku (0.35)		gemi_advanced (0.65)	
	coptic_translator (0.30)		claude_haiku (0.65)	
rouge	claude_opus (0.44)	0.25	claude_opus (0.65)	0.59
	claude_haiku (0.37)		gpt_4o (0.64)	
	coptic_translator (0.34)		claude_haiku (0.61)	

Table 1: Raw metric results

ROUGE (optimised to evaluate machine summarization) evaluate the degree of word and phrase overlap. Together, they provide a multifaceted view of translation quality. While more recent and sophisticated metrics like METEOR and ROUGE may be better markers, metrics on the other side of the spectrum behave in a more straightforward way and thus represent a bridge from human qualitative evaluation to more complex metrics.

2.2.3 Results (Coptic Texts)⁸

The results of comparing nine preprocessed⁹ translations made by different LLMs across four Coptic texts are illustrated in Table 1 and Figure 2. Each graph represents one of the Coptic texts, with the x-axis showing different evaluation metrics and the

y-axis displaying scaled, directionally normalised values of these metrics.

For texts TM 874362 and TM 874363, the specialized Coptic Translator model and the general-purpose Claude Opus performed consistently well, achieving scores near the top across most metrics. Claude Haiku also showed relatively high performance but lagged slightly behind the top performers. The Gemini and Gemini Advanced models exhibited lower performance, with scores dropping significantly in certain metrics, particularly TER and ROUGE. Other models, such as GPT variants and Claude Sonnet, displayed mixed results, performing well in some metrics and poorly in others. A similar trend was observed for TM 874364, where Claude Opus, Claude Haiku and the Coptic Translator emerged as strong performers. However, TM 874365, which is in a poor preservation state, posed challenges for all models, leading to generally lower scores and greater variability. The Gemini models and some GPT variants, in particular, struggled significantly with this text, indicating difficulty in handling degraded source material. The specialized Coptic Translator model and the general-purpose Claude Opus and Haiku achieved

⁸The code used to obtain these results can be found at <https://github.com/somiyagawa/GreekCopticMTEval>.

⁹Given the low number of texts involved in this pilot study, we opted for a semi-manual normalisation and preprocessing of the strings involved to make the metrics more meaningful. In addition to common steps like lowercase punctuation removal, we also made more case specific choices. For example, some archaic expressions were modernised (“thou art” → “you are”) and the spelling of names was uniformized (-os/-us ending). The complete steps involved will be made available together with all other results in a jupyter notebook upon acceptance of this paper.

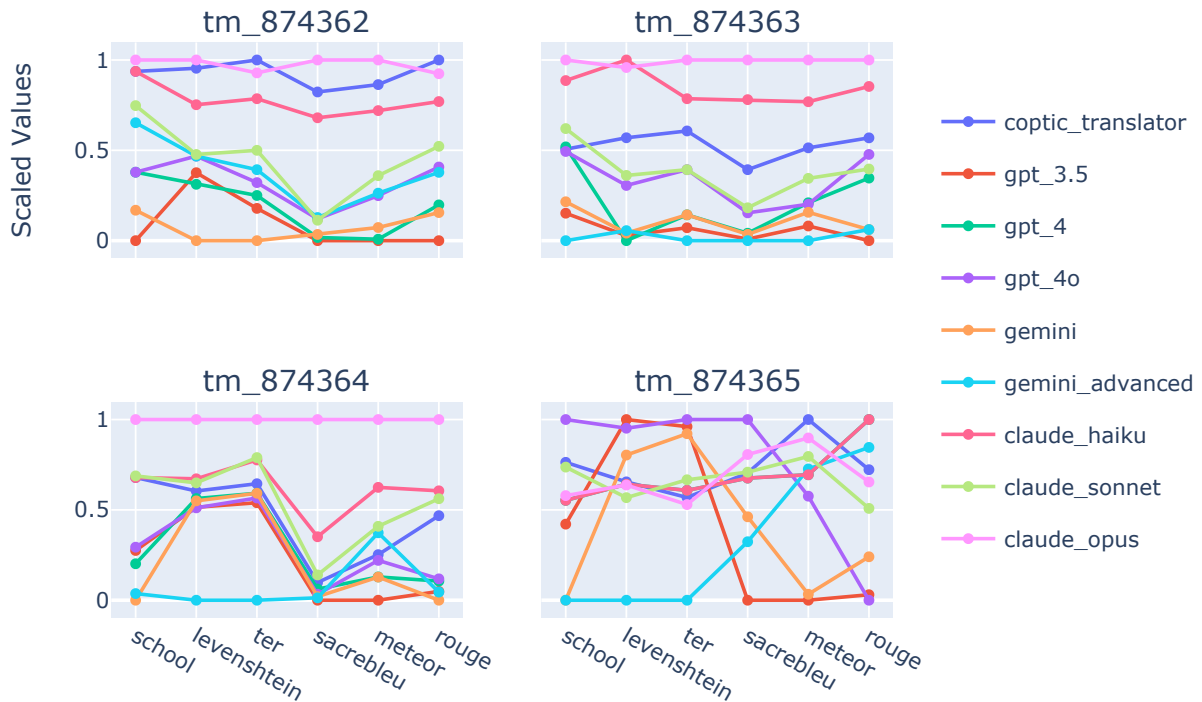


Figure 2: Scaled metrics evaluating LLMs' MT (Coptic texts)

the strongest results across the four texts. Claude Opus performed best, obtaining the highest scores on two out of four texts. In contrast, the GPT and Gemini model families struggled to produce viable translations, often generating largely irrelevant or incoherent text. When comparing across the four texts, all models found TM 874365 the most difficult, likely due to its poor preservation state. The texts in good condition, TM 874362 and TM 874363, yielded the best translation quality overall. This highlights the significant impact that the physical deterioration of source material can have on the MT process.

Examining the different metrics, we observe reasonable agreement in model rankings between “school”, Levenshtein, TER, and METEOR. However, ROUGE scores exhibit more variability, suggesting that n-gram recall may be less reliable for ancient languages. In summary, the Coptic Translator, Claude Haiku, and Claude Opus demonstrate the potential for usable MT of Coptic texts, although challenges remain with heavily damaged ostraca. The GPT and Gemini models appear unsuitable for this domain based on their inability to generate meaningful translations. GPT and Gemini tend to output Biblical quotations triggered by a proper name in the text. In the next section, we turn our attention to Ancient Greek to explore whether these findings generalize to another historically sig-

nificant language.

3 Assessing Ancient Greek-English MT

3.1 Four Greek Texts

To compare the MT metrics obtained with the four Coptic letters on ostraca, we selected four Greek texts of similar length that are also letters on ostraca: TM 817896, 89219, 89224 and 42504. For each of them, an openly available English translation made by human experts is provided on Papyri.info.

- TM 817897/ Pap.Lugd.Bat. 23 S. 7: 9 x 11.3 cm. II CE, Thebes. It is clearly a business letter, but since it is broken away at the right, the exact transaction between sender and addressee is not completely clear.
- TM 89219/ O.Ber. II 193: 9.3 x 8.7 cm. I CE, Berenike. is also a fragmentary business letter. Its end is missing, the opening and the first requests are preserved.
- TM 89224/ O.Ber. II 198: 11.5 x 9.5 cm. I CE, Berenike. From the same historical context as TM 89219, it seems to be complete. It also discusses one business-like matter together with more social elements.

- TM 42504/ O.Mich. I 91: 12.9 x 12.7 cm. III CE, Arsinoites. It is arguably in the best preservation state. This letter seems to have been sent mainly in order to obtain the favor of using borrowed oxen for an extended period of time.

3.2 Results (Ancient Greek Texts)

Notably, there seem to exist no equivalent of the CT for Ancient Greek-English translation yet. Table 1 and Figure 3 shows the metrics of the translations by the other eight LLMs. Looking at the scaled metrics across the four chosen texts, we note the following: Claude Opus, the best generic LLM in Coptic-English translation, seems to remain competitive in Greek-English translation too. Notably, the two other Claude models, Sonnet and Haiku, also perform well here. The latest model GPT-4o makes the most visible relative improvement and distances itself visibly from its predecessors GPT-3.5 (the worst performing model) and GPT-4, except in TM 817896. Upon inspection of the raw values, this text has been translated comparatively similarly across all models. We suspect this is due to its rather basic vocabulary and syntax. While Gemini Advanced scores better overall than its base model, the difference is not as pronounced, similar to the three Claude models but with slightly worse results.

From a Qualitative point of view, all models have performed better than expected, the amount of hallucinations differed radically from the one observed in translations of Coptic. In the next section, we explore whether unscaled quantitative metrics confirm this impression.

4 Comparing Results (Coptic and Greek)

Figure 4 shows the mean performance all surveyed LLMs achieved on one given metric at text level. It reveals that all MTs of Ancient Greek texts obtained better scores than their Coptic equivalents on those metrics. We note that the gap is even bigger in more complex, possibly more meaningful metrics like SacreBLEU and METEOR.

We explain this perceived difficulty of current LLMs to translate Coptic texts at the level of Ancient Greek texts with the likely very large gap in available training corpus for each corpus. This cannot be said in certainty for LLMs that are not open source, but a recent digital contribution estimated the overall size of Coptic digital papyrology

to 102,080 words across 1,973 texts. This represents about two percent of the Greek equivalent, 4,926,263 words across 58,975 texts (Riaño Ruffilanchas, 2024). Otherwise, no likely secondary factors have been found that contribute to the difference in MT performance. The length or completeness of the text appears unlikely to play a role in the eight chosen texts. (Cf. Riaño Ruffilanchas, 2024)

5 Future Directions

This study examines Coptic/Greek-English MT by LLMs in only eight texts and could be scaled and improved in multiple directions.

- Latin and other ancient languages could be included in the evaluation process. Doing so could reveal nuances in translation performance beyond mere training corpus size.
- While human translations were used as ground truth, we acknowledge that experts vary in their translating styles and preferences. Future studies might benefit from gathering translations from the same human expert en masse.
- The MTs were retrieved by accessing the Chat UIs respective to each model family and inputting a basic prompt (for example: “translate this text” + target string). Developing a more complex prompting pipeline might improve performance.
- Similarly, accessing the LLMs via API rather than via a Chat UI would allow to fine-tune the models and could alter the overall performance. This would also allow to collect large samples of translations for one individual ancient text and to better study variance within a given model.
- Experiments conducted during research for this paper suggest that the directionality of the translation greatly affects performance. English-Ancient Greek translation beared comparatively worse results than its counterpart. Many LLMs refused to perform English-Coptic translation altogether, or produced strong hallucinations with close to no grounding in the original prompt. This anecdotal experience suggests that back translation is a promising task to evaluate in the future, albeit current models will likely yield poor results in this task.

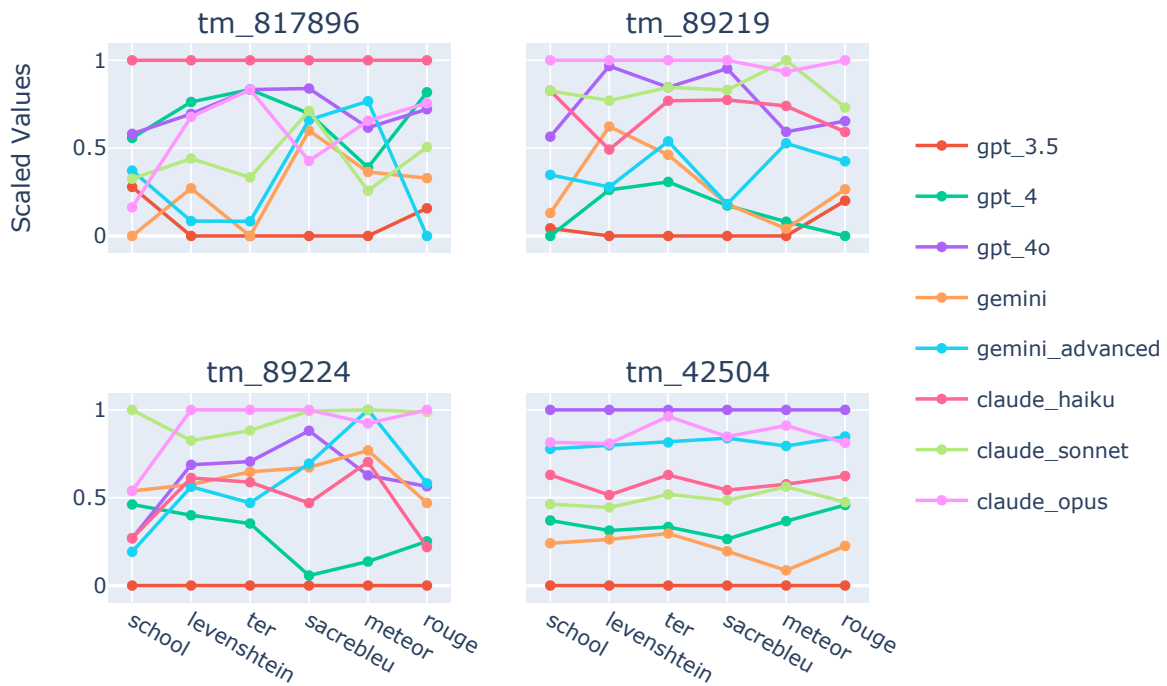


Figure 3: Scaled metrics evaluating LLMs' MT (Ancient Greek texts)

In sum, many steps could be taken to improve either the MTs themselves or their critical evaluation. However, we note that performance related considerations are not the sole concern of the scholarly assessment of MT.

- In this pilot study, paid models did not overwhelmingly outperform their free equivalents. However, this may be specific to the task as we defined it. In the face of the rapid changes in the AI industry, it can not be excluded that this will change.
- We advocate for a rethinking of translations in digital datasets of ancient texts in the LLMs era. Despite their high quality, translations by human experts are limited by the availability of specialists and not systematically added to open source databases due to a lag in publishing practices.
- There is a need for a deeper reflection upon the shared and distinct goals in MT between the industry and academia. For example, output speed is a metric current models are competing over while it bears close to no significance in the context of academic MT, where translation quality is preferred over all other aspects. Altered behaviour in translating sensitive contents is another aspect where academia and industry might have unaligned wants.

6 Conclusion

The results presented in this paper serve as recommendations for leveraging currently available Large Language Models for the Machine Translation of Ancient Greek and Coptic texts. Egyptologists and Classicists seeking to focus on just one of the nine models evaluated will find the specialised Coptic Translator or the Claude model family most beneficial. Although the best-performing version, Claude Opus, requires a paid subscription, the lighter Haiku and Sonnet models produced nearly equivalent results for Ancient Greek. The success of the Coptic Translator, a smaller specialised tool fine-tuned from a larger model, stands out in the context of increasingly capable general-purpose models. GPT-4o, the newest model included in this pilot study, did not show significant improvement for Coptic but did for Ancient Greek. Aside from GPT-3.5, which consistently underperformed, the other three models form a middle tier in performance. However, scholars are not limited to selecting only one model, especially given the rapid development and deployment of new models, alongside related ethical considerations. We advocate for the joint use of multiple models to provide the best support for human experts.

Clearly, more work is needed to fully understand how scholars can benefit from LLM-powered translations. Nevertheless, it is evident that multilin-

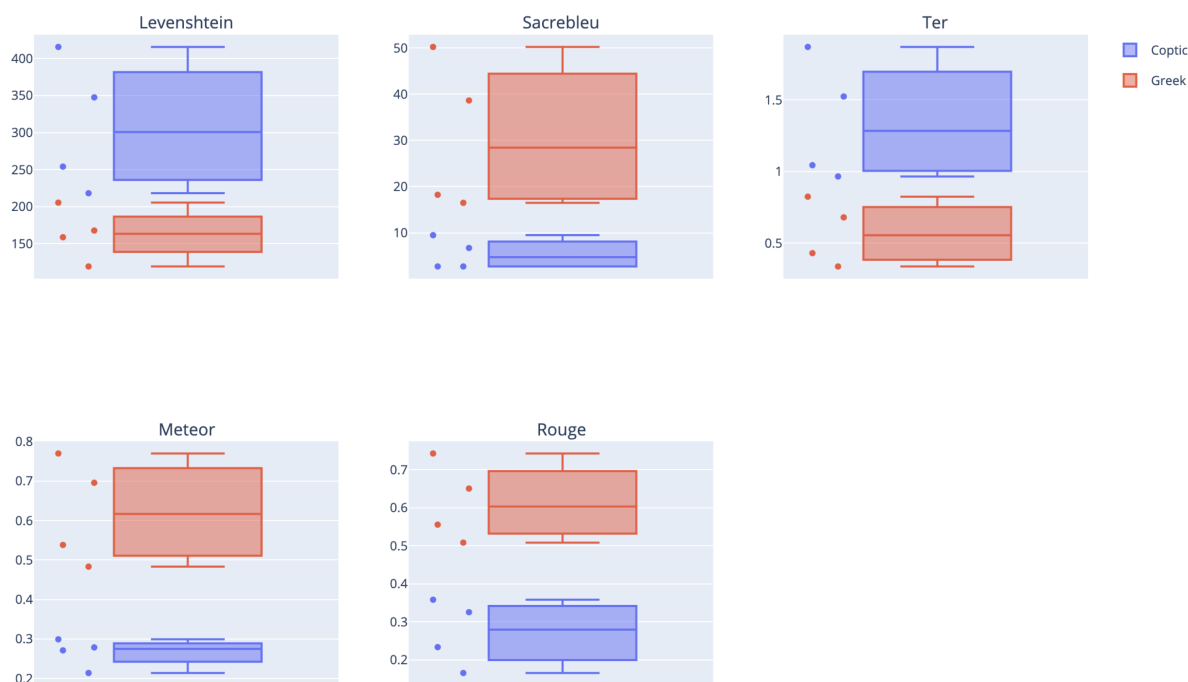


Figure 4: Distribution of mean raw metrics

gual Digital Papyrology will benefit from the semi-automatic generation of metadata (e.g., translations, summaries) enabled by these technologies. The potential to link and integrate previously monolingual datasets seems to outweigh the risks of hallucination, which can be formally addressed using NLP metrics like those employed in this study.

Limitations

One of the main limitations to scaling up the use of large language models (LLMs) or their fine-tuning for specific tasks is the considerable cost associated with these processes. The financial burden includes the expense of the high-performance GPUs required for learning and inference, as well as the cost of access to state-of-the-art models, often hidden behind paywalls. In addition, access to the user interfaces of several advanced LLMs usually entails additional costs, making comprehensive evaluations across multiple models a financial challenge. These barriers can prevent researchers and smaller institutions from fully exploiting the possibilities offered by LLMs, potentially limiting the diversity and breadth of research in this field.

Ethics Statement

While our discussion focuses on improvements in Machine Translation (MT), we emphasize the importance of supporting human-human collaboration

in scientific undertakings. The development of MT should not overshadow the important role of human translators and experts in the translation process. We advocate not only a “human-in-the-loop” approach, where human oversight and collaboration are essential to ensure the accuracy and reliability of translation. We also stress that a responsible and ethical MT technology must be human-centered.

References

- Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- Josephine Balmer. 2009. Jumping their bones: Translating, transgressing and creating. *Living Classics: Greece and Rome in Contemporary Poetry in English*, pages 43–64.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

- Anne Boud'hors and Esther Garel. 2019. [Ten Coptic Ostraca at the IFAO](#). *Bulletin de l'Institut français d'archéologie orientale (BIFAO)*, (119):51–77.
- Peter Burke. 2017. *Popular culture in early modern Europe*. Routledge.
- Sarah J Clackson. 2004. Papyrology and the utilization of coptic sources. In *Papyrology and the history of early Islamic Egypt*, pages 21–44. Brill.
- Sonja Dahlgren. 2018. Outcome of language contact: Transfer of Egyptian phonological features onto Greek in Graeco-Roman Egypt. *Journal of Greek Linguistics*, 18(1):155–165. Publisher: Brill.
- Maxim Enis and Andrew Megalaa. 2024. [Ancient voices, modern technology: low-resource neural machine translation for coptic texts](#).
- Federico Gaspari. 2024. The History of Translation Technologies. In *The Routledge Handbook of the History of Translation Studies*. Routledge. Num Pages: 15.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81. Barcelona, Spain. Association for Computational Linguistics.
- Dean P. Lockwood. 1918. [Two Thousand Years of Latin Translation from the Greek](#). *Transactions and Proceedings of the American Philological Association*, 49:115–129. Publisher: [Johns Hopkins University Press, American Philological Association].
- Outi Merisalo. 2015. [Translating the Classics into the vernacular in sixteenth-century Italy](#). *Renaissance Studies*, 29(1).
- Holger Nørgaard. 1958. [Translations of the Classics into English before 1600](#). *The Review of English Studies*, 9(34):164–172. Publisher: Oxford University Press.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. [Yara parser: A fast and accurate dependency parser](#). *Computing Research Repository*, arXiv:1503.06733. Version 2.
- Nicola Reggiani. 2017. *Digital Papyrology I*. De Gruyter, Berlin, Boston.
- Daniel Riaño Rupilanchas. 2024. [Counting the number of words in greek and latin papyri](#).
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Peter van Minnen. 1993. [The Century of Papyrology \(1892-1992\)](#). *The Bulletin of the American Society of Papyrologists*, 30(1/2):5–18. Publisher: American Society of Papyrologists.
- Marja Vierros and Erik Henriksson. 2017. Preprocessing Greek Papyri for linguistic annotation. *Journal of Data Mining & Digital Humanities*, (Towards a Digital Ecosystem: NLP. Corpus infrastructure. Methods for Retrieving Texts and Computing Text Similarities). Publisher: Episciences. org.
- Jennifer Westerfeld. 2016. [Decipherment and Translation: An Egyptological Perspective](#). *CR: The New Centennial Review*, 16(1):29–36. Publisher: Michigan State University Press.
- Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. [Google's neural machine translation system: Bridging the gap between human and machine translation](#). *ArXiv*, abs/1609.08144.
- Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. [Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond](#). *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

The Social Lives of Literary Characters: Combining citizen science and language models to understand narrative social networks

Andrew Piper[†] Michael Xu[†] Derek Ruths[‡]

[†]Department of Languages, Literatures, and Cultures [‡]School of Computer Science
McGill University

Abstract

Characters and their interactions are central to the fabric of narratives, playing a crucial role in developing readers' social cognition. In this paper, we introduce a novel annotation framework that distinguishes between five types of character interactions, including bilateral and unilateral classifications. Leveraging the crowd-sourcing framework of citizen science, we collect a large dataset of manual annotations (N=13,395). Using this data, we explore how genre and audience factors influence social network structures in a sample of contemporary books. Our findings demonstrate that fictional narratives tend to favor more embodied interactions and exhibit denser and less modular social networks. Our work not only enhances the understanding of narrative social networks but also showcases the potential of integrating citizen science with NLP methodologies for large-scale narrative analysis.

1 Introduction

Characters and their interactions are a fundamental feature of storytelling. As a prominent dimension of cognitive literary theory has argued, fictional characters provide readers with the opportunity to identify with other imaginary human beings and model social relationships (Zunshine, 2006; Mar et al., 2006, 2009, 2010; Palmer, 2004). According to this theory, the interactions among characters and the resulting social networks provide an important training ground for the development of social cognition (a.k.a. Theory of Mind) (Kidd et al., 2016).

Well over a decade ago, a robust body of work began to emerge in NLP to address the extraction of social networks from narrative texts (Elson et al., 2010; Lee and Yeung, 2012; Agarwal et al., 2013, 2014). That work established important methodological foundations for the study of literary social networks, which were understood to be a key

component of research in the Digital Humanities (Moretti, 2011).

Nevertheless, this work was faced with two significant challenges: 1) automated methods never exhibited robust levels of accuracy to be applicable to real-world cases; 2) the costliness of deriving high-quality training data made it difficult to build more accurate models.

In this paper, we aim to address these challenges to further the study of narrative social networks by focusing on the following methodological contributions:

1. Establishing a novel annotation framework that includes five distinct interaction types, including a second level classification of *bilateral* versus *unilateral* types (i.e. whether one of the characters involved is aware of the interaction).

2. Illustrating the value citizen science can have for research in NLP by releasing and validating a large-scale dataset of manual annotations of character interactions (N=13,395).

3. Training, validating, and publicly sharing a performant, open-weight small language model or SLM (Phi3-7B) for the task of interaction labeling (F1 = 0.70).

4. Testing the effects that different genres and audience types have on social network structure within a sample of contemporary books (N=390). Here we demonstrate a proof-of-concept analysis of the ways in which social networks in stories are shaped by genre and audience factors such as fictionality, cultural context, prestige, and expected reader age.

We conclude with a discussion of areas for future work.

2 Prior Work

Some of the earliest theoretical work concerning the value of social networks for literature was undertaken by Moretti (2011) and Woloch (2009). Woloch (2009)'s theory of *character-space* has

been particularly influential. This theory highlights the skewed distribution of attention around a primary character, referred to as ‘the one and the many’ structure by Woloch. This work established an important theoretical framework for studying character *relations* in addition to individual characters (Propp, 1968; Frow, 2014).

Early methodological work on the extraction of character networks was undertaken by several researchers (Sudhakar and Cristianini, 2013; Agarwal et al., 2014; Trovati and Brady, 2014; Nijila and Kala, 2018). Much of this work focused on the use of sentence-level co-occurrence or subject-verb-object triplets as the foundation of building character interactions. Labatut and Bost (2019) provide an extensive survey of methods of social network extraction applied to cultural texts (i.e. stories and screenplays), ranging from interaction identification to network analysis methods.

In terms of applications, Mac Carron and Kenna (2012) analyzed character networks within three European epics (*The Iliad*, *Beowulf*, *Tain Bo Cuil-lange*) to understand their relation to contemporary real-world social networks. Volker and Smeets (2020) compared fictional networks in Dutch literature with real-world networks with respect to racial groups. And Ardanuy and Sporleder (2014) and Agarwal et al. (2021) used social networks as a mechanism to detect book genres.

Dialogue networks have also been studied as a subset of literary social networks (Elson et al., 2010; Waumans et al., 2015), with a similar principal applied to the study of drama (Algee-Hewitt, 2017; Lee and Lee, 2017; Fischer and Skorinkin). Finally, substantial work has focused on the detection of *relationship* types (instead of individual interactions), such as kinship ties (Iyyer et al., 2016; Chaturvedi et al., 2016; Massey et al., 2015; Makazhanov et al., 2014) and conflict groups (Smeets et al., 2021).

3 Methods

3.1 Defining Character Interactions

We define a character interaction as occurring when *a character / group of characters engage in an action that involves another character / group of characters within the story world of a narrative*. This definition allows for the inclusion of a single character or group at the level of the agent or patient (object). And following the work of Agarwal and Rambow (2010), it also supports a base-level

distinction between “bilateral” and “unilateral” interactions, i.e. when both characters are aware of the interaction or only one of the characters is aware. While an interaction requires two characters / groups to be an interaction, it does not require cognizance of the action by the patient.

Accordingly, we identify five possible types of interactions: *communicating*, *thinking about*, *observing*, *touching* (physical contact), and *associating* (which we use as a catch-all). Only *observing* and *thinking about* can be unilateral. Table 1 provides example sentences of the different interaction types.

3.2 Using Citizen Science for Manual Annotation

“Citizen science” is a term used to describe the general public engagement in scientific research activities (Consortium et al., 2013). Citizen science projects have annotated over 250 million pieces of data over the past two decades, ranging from the identification of galaxies, bird species, to the location of marine-based trash. Research shows that data produced by citizen science projects can be of high quality and correlate strongly with expert opinion when best-practices are employed (Kosmala et al., 2016; Wiggins and He, 2016). It also provides a cost-efficient means of data collection (Sauermann and Franzoni, 2015).

Several citizen science projects have emerged in the humanities in recent years (Ridge, 2016; Terras, 2015; Dobрева and Azzopardi, 2014), although the quantity of projects is still small compared to the natural sciences which represent an estimated 80% of all projects (Hecker et al., 2018). To date, most citizen science initiatives in the humanities have focused on document transcription. Our project, called *Citizen Readers*, uses the popular platform Zooniverse.org and focuses on text annotation common to the NLP community, which has traditionally been undertaken through fee-based crowd-sourcing platforms. Our project thus seeks to illustrate the opportunities that await both the humanities and NLP through the use of volunteer citizen scientists.

Figure 1 provides an illustration of our task structure. Participants were presented with two-sentence passages, where the first sentence serves as the context and the second sentence includes two highlighted characters for interaction classification. Passages were randomly sampled from the CONLIT dataset of contemporary books (Piper,

Type	Sentence
Associating	When Admiral Bloch left the Dauntless, he placed me in command of the fleet.
Communicating	Then I saw them waving at me from the far end of the restaurant.
Observing-Uni	She peers out at the sniper , but the angle is enough to hide her from his sight.
Observing-Bi	Ange looked over at me , then sprung up and headed my way.
Thinking	She thinks about Ned in his brown Doc Martens.
Touching	Dr. Fell gently brought Ethel Pusster to her feet.

Table 1: Example sentences of our various interaction types. Boldface represents the highlighted characters.

2022), which contains twelve different genres of fiction and non-fiction books, and characters were automatically detected using bookNLP (Bamman, 2021).

Participants were then asked a series of conditional questions: 1. Are the highlighted characters interacting in the story? If yes: 2. Is the interaction unilateral or bilateral? Given their answer: 3. What kind of interaction is it? for which the relevant classes were presented.

In addition to the task itself, Zooniverse provides an area for a custom tutorial, a field-guide with more in-depth descriptions, an about page to inform participants about the goals and intentions of the project, and a talk area where moderators can respond to participant questions. For this project we employed four student moderators who were indispensable in responding to the volume of questions.

A total of 1,915 citizen scientists participated in our project completing 73,648 unique annotations. The project took three months to complete. Out of the initial 19,006 passages posted to the platform, 15,641 were annotated by three or more annotators. The total number of passages where we observed a majority consensus on the label was 13,395. We found that 1,189 participants (or 62%) annotated five or more passages, and only 249 annotated a single passage. We also observe the Pareto principle at work, with 20% of our participants completing 72% of our annotations.

In order to assess the quality of annotations by citizen scientists, we hired three trained students to annotate a small subset of passages (N=261). We then compared agreement scores for three cases: inter-student annotations, inter-citizen scientist annotations, and student-citizen scientist annotations. We calculate Fleiss’s Kappa scores for two conditions: all annotations and only those with majority votes. As we can see in Table 2, student annotators exhibit slightly higher agreement for all annota-

tions but when conditioning on those with majority agreement the scores converge. We also show very high levels of agreement between student and citizen-scientist majority annotations, suggesting the high quality of our final annotations.

Condition	student	citizen scientist	student-citizen
All annotations	0.48	0.41	0.49
Majority votes	0.50	0.51	0.79

Table 2: Comparison of agreement scores using Fleiss’s Kappa between students and citizen scientists.

We present the distribution of interaction types in our 13,395 majority-labeled passages in Fig. 2. The most common label is “no interaction,” followed by “communicating” and “associating.” While the other three types are far rarer, we will see in later sections their significance. We release our data set, known as the “Citizen Readers for Character Interactions” dataset (*CR4Interact*), which is available in our project’s long-term repository.¹

3.3 Finetuning an SLM for Interaction Detection

3.3.1 Training and Test Data

We then use our labeled data to fine-tune and test the performance of a small-language model (SLM) for the task of interaction type detection. For training and testing purposes, we extracted an equal amount of data from each class and a confidence score to partition the data, understood as the average agreement percentage per passage. We use passages with the highest confidence to build the test dataset, moderate confidence for the validation dataset, and the remaining data for the training dataset, as shown in Table 3. This approach guarantees the most accurate evaluation results possible, although it introduces some difficulties for the SLM because it will be trained on the lowest quality data.

¹<https://doi.org/10.5683/SP3/QMIARS>

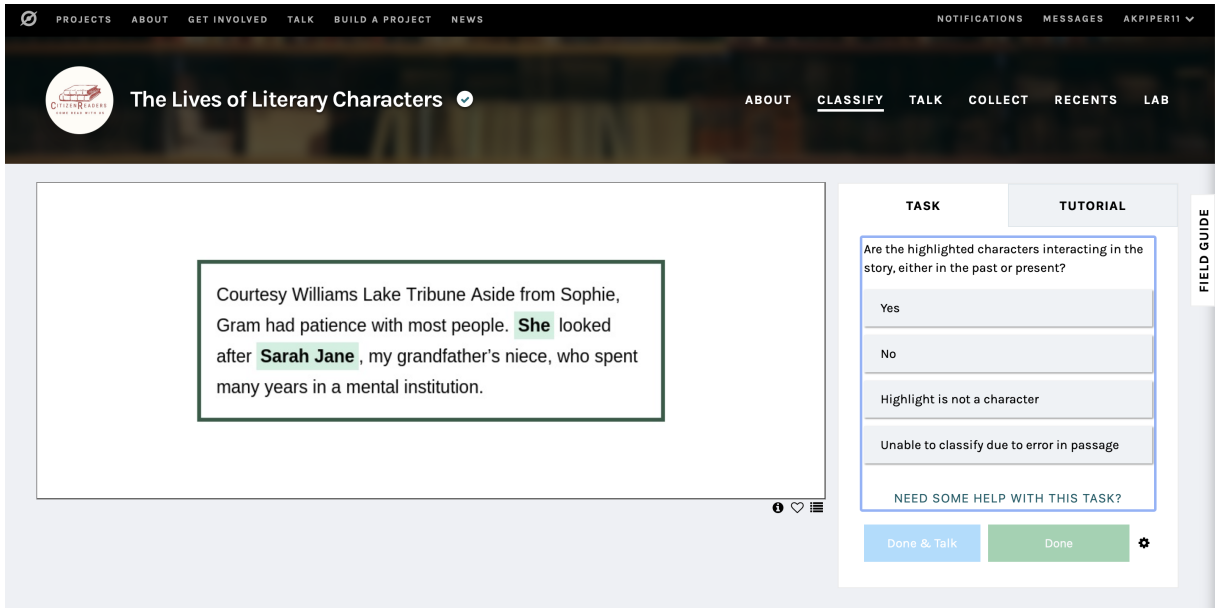


Figure 1: Image of our annotation task on Zooniverse.

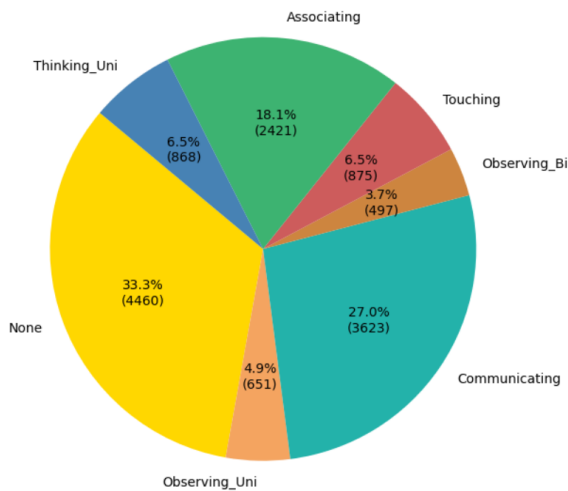


Figure 2: Label distribution in the annotated data.

3.3.2 Model Selection

To select the optimal combination of models and prompting strategies, we compare Phi-3-Mini-4K-Instruct, Phi-3-Small-8K-Instruct, Mistral-7B-Instruct-v0.3, Llama-3-8B-Instruct with GPT-4o (version 2024-05-13). We use the different prompting strategies listed in Table 4 and fully described in the appendix A and use a small balanced subset of our test data consisting of 10 passages per category for initial testing.

Among the open-source models, the detailed prompt on the Phi-3 (7B) achieves the highest accuracy of 0.73, as shown in Table 4. Phi-3 (7B) and GPT-4 are the only models that surpass 0.70 accu-

Class	Training	Validation	Test
Associating	2032	225	225
No	4085	225	225
Thinking	443	225	225
Communicating	3223	225	225
Observing	721	225	225
Touching	437	225	225
Total Amount	10941	1350	1350

Table 3: Class distribution and total amounts for the training, validation, and test datasets.

racy, and we anticipate an additional performance improvement after the finetuning stage. Note these performance numbers are based on only a small subsample of our test data.

Model	Base	Detailed	One-shot	Many-shot
Phi-3 (3.8B)	0.30	0.62	0.48	0.30 (8-shot)
Phi-3 (7B)	0.55	0.73	0.56	0.55 (8-shot)
Mistral (7B)	0.43	0.58	0.22	0.17 (8-shot)
Llama-3 (8B)	0.52	0.47	0.58	0.45 (5-shot)
GPT-4	0.65	0.70	0.73	0.72 (10-shot)
Tuned Phi-3	0.68	0.80	0.78	0.77 (8-shot)

Table 4: Performance comparison across different models and prompt strategies using the initial small, down-sampled test dataset. Tuned Phi-3 (7B)’s performance is included for reference.

3.3.3 Our model: Phi-3-interact

We use a single A100 PCIE 80GB to finetune the Phi-3 (7B) on the training dataset and monitor the loss on the validation dataset. We set the learning

rate to $2.5e-5$ and utilize the `paged_adamw_8bit` optimizer to reduce memory usage and accelerate tuning speed.²

After finetuning on the 10K training data, the model’s accuracy increases from 0.58 to 0.71 on the validation dataset. Subsequently, tuning the model on the smaller, but higher-quality 1K validation dataset further improves accuracy to 0.80 on the small test dataset (as shown in Table 4) and from 0.727 to 0.735 using the full test dataset, surpassing GPT-4’s 0.70 accuracy on the same data.

Table 5 indicates that both models struggle to distinguish the “No interaction” class, which is most often confused with the “Associating” class. Otherwise, categories range from a low of 0.70 accuracy (Thinking) to a high of 0.96 (Touching). A closer examination of the classification errors reveals that both models tend to assign positive interactions for hypothetical scenarios such as “If I could talk with her,” which annotators were instructed not to consider as interactions because they do not actually take place in the storyworld. Our finetuned model, Phi-3-interact, demonstrates significantly higher precision in the “No interaction” category compared to GPT-4, indicating that Phi-3-interact is more reliable for predicting the absence of interaction.

Class	Precision	Recall	F1	Acc
Associating	0.75	0.56	0.64	0.56
	0.59	0.75	0.66	0.75
Communication	0.75	0.89	0.81	0.89
	0.64	0.85	0.73	0.85
No	0.89	0.21	0.34	0.21
	0.54	0.20	0.29	0.20
Observing	0.82	0.94	0.88	0.94
	0.89	0.79	0.84	0.79
Thinking	0.60	0.85	0.70	0.85
	0.61	0.71	0.65	0.71
Touching	0.76	0.96	0.85	0.96
	0.92	0.90	0.91	0.90
Mean	0.76	0.74	0.70	0.74
	0.70	0.70	0.68	0.70

Table 5: Performance metrics using our fine-tuned model (Phi-3-interact, upper row) and GPT-4 (lower row) on the full test dataset.

3.4 Constructing social networks from our data

In order to analyze social networks at the book level, we first sub-sample the CONLIT dataset

²<https://huggingface.co/ChunB1/Phi-3-interact>

down to 390 books to represent the genre and audience categories described more fully in Section 4 and shown in Table 6. We use bookNLP (Bamman, 2021) to perform sentence tokenization, entity recognition (NER tag “PER”), and co-reference resolution on the book level data. From there we extract all possible candidate pairs of characters for every sentence in each book for a total of 3,928,602 possible interactions. We then use our fine-tuned Phi-3 model to label all interactions.

To construct the nodes of our networks, we use the master character IDs provided by bookNLP that are derived from the co-reference resolution step. This gives us a list of unique characters per book. We then construct weighted edge lists for each book, where an edge represents the sum of all interactions between two characters. We then construct network graphs for each book for all interaction types and one aggregate network per book.

Finally, given our edge lists we then extract the following set of network statistics for each book according to two conditions: all characters and only characters whose degree (number of relationships) is five or greater, in order to focus on more significant characters.

Protagonist Centrality. The degree of the most connected character, normalized by dividing by the total number of edges. Equivalent to the percentage of all relationships consumed by the most central character.

Density. The ratio of the number of actual edges in a graph to the potential number of edges. Ranges from 0 to 1.

Transitivity. The global transitivity of the graph also known as the clustering coefficient. This measures the ratio of the number of closed triplets (or triangles) to the total number of triplets (both open and closed) in the network.

Average Shortest Path. The average length of the shortest path between all pairs of nodes in the network.

Modularity. Measures the strength of division of a network into communities, quantifying the degree to which nodes within the same community are more densely connected to each other than to nodes in different communities. Higher modularity values indicate stronger sub-community structure. Here we use the Fast and Greedy algorithm.

4 Analysis

In this section, we aim to illustrate the potential utility of our data for the large-scale study of cultures of storytelling. We measure the effects of the following four stylistic and audience categories on the distribution of character interaction types and the resulting book-level social networks:

Fictionality. Here we test for the effects that fiction has on social interactions. For interactions we look at all adult genres. For social networks, we sample books from the Prizewinners and Best-sellers categories for fiction and Biography for non-fiction.

World. We test for the effects that books published in non-Western cultures have on social interactions. Specifically we look at books published in English in India, South Africa, and Nigeria that were reviewed in major literary reviews in their respective regions and compare them to our Western Prizewinner category.

Prestige. Prior work has identified strong stylistic differences between best-selling fiction and fiction written to appeal to literary elites on prize committees (Piper and Portelance, 2016). Here we test whether these findings extend to social interactions and their resulting networks.

Youth. For this category we compare middle-school fiction with adult fiction as represented by Prizewinners and Bestsellers. We expect to observe strong effects that are designed to make narratives more accessible to younger readers.

4.1 Book type effects on character interactions

In order to study the effects of our book categories on interaction types, we utilize count data to compute the log odds ratios through Fisher’s Exact Test, focusing on the rate of each interaction type relative to the overall interaction rate for each category. Our findings indicate that fiction uniquely exhibits statistically significant effects (Fig. 3).

Specifically, non-fiction surpasses expectations in rates of communication and association, whereas fiction emphasizes observation and physical contact. These results corroborate existing theories that highlight the importance of embodied behavior in fictional narratives (Caracciolo and Kukkonen, 2021; Piper, 2024). Interestingly, there is no meaningful difference in the rate of unilateral versus bilateral actions in either corpus. Fiction does not indicate a preference for unilateral interactions as

might be hypothesized by the strong emphasis on social cognition theories of reading fiction (Zunshine, 2006).

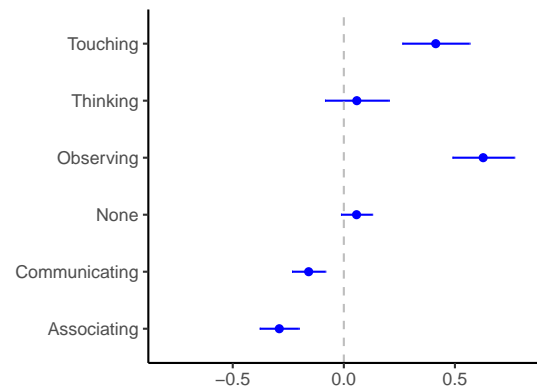


Figure 3: Log odds of interaction type appearing in fiction compared to non-fiction.

4.2 Book type effects on social networks

The first notable fact about our book-level social networks is the long tail of degree distribution (i.e. the number of relationships per character). For both fiction and non-fiction more than 86% of characters per book have fewer than five relationships in their respective social networks, suggesting a long tail of minor characters and a small, central core of main characters (Woloch, 2009).

Table 6 shows the results of our regression analysis for the aggregate social network structures in our sampled books by category. For our analysis, we require five or greater connections for inclusion in the network given the long tail of very minor characters (though we note that the overall results do not change in any meaningful way if we include all characters).

Our analysis reveals that fictional narratives exhibit significantly increased density and transitivity, coupled with lower modularity, average shortest path, and time to completion. To quantify the effect sizes, we converted these findings into Cohen’s d values. This translation demonstrates substantial effects, with values ranging from $d=1.3$ for density to $d=-2.2$ for modularity, indicating large to very large effects. We also find that non-fictional narratives take 50% longer to complete their social networks when compared to fictional narratives.

These results indicate how strongly fictional narratives tend to create denser, more connected networks than biographical narratives. Although biographies condition explicitly on a single life, they

exhibit on average more modular narrative structures (strongly connected components that are less connected overall). Fiction by contrast maintains a denser overall relational network, one in which the introduction of significant new characters is completed much earlier than in biographical narratives.

Category	Fictionality	World	Prestige	Youth
Density	+ (***)	.	.	+ (***)
Centrality	.	.	.	+ (*)
Transitivity	+ (***)	.	.	+ (***)
Completion	- (***)	.	.	.
Modularity	- (***)	.	.	- (***)
Shortest Path	- (***)	.	.	- (***)

Table 6: Results of the regression analysis. +/- refer to positive or negative effects and the number of asterisks refer to p-value magnitude (* < 0.01, ** < 0.001, *** < 0.0001). A period denotes no meaningful effect.

These results lend further support to prior work suggesting a “small world effect” of fictional narratives. Prior work has shown that fictional characters cover smaller geographic distances (Matthew Wilkens, *forthcoming*) and that fictional narratives exhibit considerably lower overall informational surprise, favoring narrative ‘exploitation’ (covering familiar characters and situations) over narrative ‘exploration’ (introducing new characters, themes, and situations) (Piper et al., 2023). Here we can add the denser social network structures as a further index of this small world effect of fictional narratives.

Within fiction, we observe no meaningful effects for either social prestige (books receiving literary prizes) or books published in non-Western cultures. Youth books on the other hand exhibit very clear signals of simplified social networks with lower modularity, shorter paths, more centralized protagonists and greater relationship density. Youth books in other words tend to amplify the effects of fictionality.

Our results suggest two important points: the first is that the expected values we are observing with respect to major distinctions like fiction/non-fiction and adult/youth indicate that our social networks are capturing important information about the underlying social structures of the sampled books. While we do not yet have a way to validate the accuracy of the constructed social networks from local character interactions these results give us confidence that broadly speaking we are capturing meaningful differences in narrative construction. That said, the more subtle differences

we observe between different cultural contexts or levels of social prestige may yet be due to measurement error. Future work will want to investigate this more fully.

One further question we investigated was whether interaction-type sub-networks differ significantly from the larger networks in which they are imbedded. Do observational or communicative or other types of interactions lead to structurally different properties that might initiate new theories about the relationship between social interactions and social networks within narratives?

To measure structural equivalence between social networks, we utilized cosine similarity as our primary metric. Structural equivalence traditionally involves assessing the commonality of neighbors between pairs of vertices; however, a simple count of common neighbors does not account for variations in vertex degrees or the broader distribution of common neighbors among other vertex pairs. Cosine similarity addresses these limitations by considering the degree of the vertices and their neighbors.

In our method, we treat the rows/columns of the adjacency matrix as vectors. The cosine similarity between two vertices i and j is calculated as the cosine of the angle between their corresponding vectors. Mathematically, the cosine similarity of vertices i and j is defined as the number of common neighbors divided by the geometric mean of their degrees. This measure produces a value ranging from 0 to 1, where 1 indicates that the two vertices share exactly the same neighbors, and 0 indicates no common neighbors. For vertices with a zero degree, we conventionally set their cosine similarity to 0.

When doing so, interestingly we find no meaningful distinctions between the structural similarity of different types of sub-graphs when compared to the main graphs to which they belonged (Fig. 4). While the rate of different interaction types differed strongly between fictional and non-fictional narratives, for example, the underlying network structures to which they contribute do not.

5 Conclusion

Understanding narratives at large scale is a core concern of the Digital Humanities (Underwood, 2019; Piper, 2018). The social interactions of characters (Zunshine, 2006) and the resulting social networks (Moretti, 2011; Woloch, 2009) have long

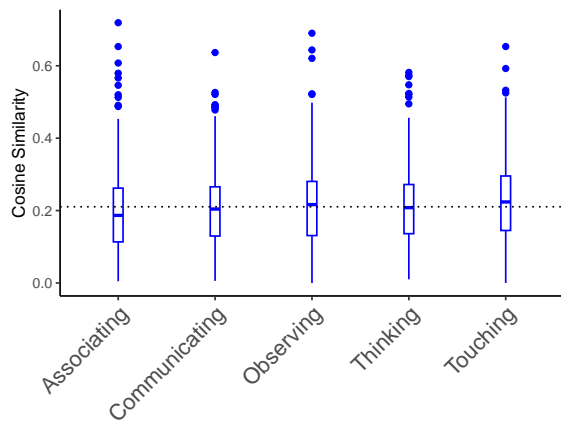


Figure 4: Comparison of structural similarity of subgraphs to main graphs by type. The dotted line represents the global mean.

been theorized as important dimensions of human storytelling. In this paper we have endeavored to illustrate the potential that citizen science has as a means of generating data for training and testing language models towards the goal of understanding the social lives of characters.

With respect to Citizen Science as a mechanism of crowd-sourced text annotation, we find annotation quality on par with trained student annotators. As prior work has suggested, Citizen Science projects achieve the same quality standards as other approaches and bring with them the affordances of a volunteer, community-based approach to scientific discovery (Kosmala et al., 2016; Wiggins and He, 2016). We hope more projects in NLP and DH will utilize this significant resource.

With respect to narrative understanding, we have identified two notable findings in our work. First, fictional narratives strongly favor embodied forms of interactions such as physical contact and sensory perception in support of prior work (Piper, 2024). Second, fictional narratives also strongly favor far denser and less modular social networks. The physicality of relations between characters is amplified by the overall *connectedness* of characters (Mar and Oatley, 2008).

We could add here one additional negative finding: social networks built around individual interaction types do not appear to differ from the overall narrative social networks to which they belong. Social networks built around contact, communication, observation, etc. follow the same patterns as the full network. So while it appears interaction types are useful for distinguishing fictional narratives,

they do not contribute much to our understanding of the larger social network structures.

We also highlight a number of areas for future work: both our SLM and GPT indicate significant limitations with respect to the detection of non-interactions, which has also been demonstrated with respect to grammatical models (Agarwal et al., 2013). Much of this can be attributed to the fuzzy boundary around the concept of “interaction” – when two characters are grammatically proximate the rejection of their interaction depends on a number of subtle factors (hypotheticality, co-presence versus interaction, etc.). Future work will want to further explore this boundary in particular.

A second key area is the validation of the social networks themselves. Book-level data on narrative social networks remains a costly endeavor. To date, the field still lacks reasonably sized ground truth when it comes to validating book-level social networks. While we show that our constructed social networks from local interactions align well with theoretical expectations, further validation of their accuracy awaits.

Finally, while we introduce a novel interaction framework in our work, future work will want to think about further nuance with respect to labeling interactions. Our work does not address the valence of interactions, an important property of narrative relations (Smeets et al., 2021), nor does it address overall relationship types (such as kinship or narrative properties such as antagonist, etc.). These too can be valuable frameworks for understanding the structural properties of narratives.

We hope that our publicly shared training data and SLM can be useful tools for researchers to further study the nature of narrative social networks.

Limitations

As we mention above, our work is subject to a number of limitations. First, we note that despite the relatively large size of our training data particular interaction types are significantly less well represented (e.g. observing, thinking, touching). Future work will want to concentrate on expanding our understanding and coverage of those categories. As we also note, while our ability to identify specific types of interactions is high, our ability to distinguish between non-interactions and associations is weak. Future work will want to explore this boundary more fully.

We also highlight that future work will want to

provide book-level annotations of social networks to validate the accuracy of moving from local interaction prediction to global social network modeling.

Another important limitation is the cultural specificity of our data. While our data is drawn from a broad array of genres and a few world cultures, they are limited to the English language. Future work will want to assess cultural differences more deeply with respect to interaction types and social networks.

Ethics Statement

Relying on crowd-sourced labor brings with it important ethical considerations, specifically around fair labor practices and the representativeness of the participating community (Harmon and Silberman, 2019). Citizen Science makes two important contributions to these issues: first, it relies on volunteer rather than paid labor and thus depends on the project-specific interest of participants. Platforms like Zooniverse further contribute to this through the use of About pages, team descriptions, and talk pages where participants can interact with researchers. Participants are far more aware of research goals of a project when it comes to Citizen Science.

In addition to promoting greater project transparency, Citizen Science projects also promote greater researcher-citizen connections, which can help support the democratization of scientific knowledge and facilitate participant learning (Bonney et al., 2016) without sacrificing quality.

We note that while Citizen Science projects can lower the cost of large-scale annotations they do require far more planning and design investment. The initial adaptation of tasks to a particular platform can take time, but we have found that after initial learning projects can take about 2-3 months to prepare for launch. Additionally, because Talk pages are actively used by participants it is essential to have moderators available to handle the volume of queries from users. Nonetheless, all of this can contribute to more transparency and involvement by citizens which is a decidedly positive contribution.

Acknowledgements

The authors wish to thank Robert Budac and Geoffrey Rockwell for their management of the Zooniverse Citizen Readers platform and the Social Sciences and Humanities Research Council of Canada

for their generous support of this research.

References

- Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014. [Frame semantic tree kernels for social network extraction from text](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–219, Gothenburg, Sweden. Association for Computational Linguistics.
- Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1202–1208.
- Apoorv Agarwal and Owen Rambow. 2010. Automatic detection and classification of social events. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1024–1034.
- Divya Agarwal, Devika Vijay, et al. 2021. Genre classification using character networks. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 216–222. IEEE.
- Mark Algee-Hewitt. 2017. Distributed character: Quantitative models of the english stage, 1550–1900. *New Literary History*, 48(4):751–782.
- Mariona Coll Ardanuy and Caroline Sporleder. 2014. Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 31–39.
- David Bamman. 2021. Booknlp. a natural language processing pipeline for books. <https://github.com/booknlp/booknlp>. Accessed: 2022-01-30.
- Rick Bonney, Tina B Phillips, Heidi L Ballard, and Jody W Enck. 2016. Can citizen science enhance public understanding of science? *Public understanding of science*, 25(1):2–16.
- Marco Caracciolo and Karin Kukkonen. 2021. *With bodies: Narrative theory and embodied cognition*. Ohio State University Press.
- Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Socientize Consortium et al. 2013. Green paper on citizen science. *Citizen Science for Europe. Towards a better society of empowered citizens and enhanced research*. Brussels.

- Milena Dobрева and Daniela Azzopardi. 2014. Citizen science in the humanities: a promise for creativity. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.
- David Elson, Nicholas Dames, and Kathleen McKeown. 2010. [Extracting social networks from literary fiction](#). In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.
- Frank Fischer and Daniil Skorinkin. Social network analysis in russian literary studies. *The Palgrave handbook of digital Russia studies*, page 517.
- John Frow. 2014. *Character and person*. Oxford University Press.
- Ellie Harmon and M Six Silberman. 2019. Rating working conditions on digital labor platforms. *Computer Supported Cooperative Work (CSCW)*, 28(5):911–960.
- Susanne Hecker, Lisa Garbe, and Aletta Bonn. 2018. *The european citizen science landscape—a snapshot*. UCL Press.
- Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.
- David Kidd, Martino Ongis, and Emanuele Castano. 2016. On literary fiction and its effects on theory of mind. *Scientific Study of Literature*, 6(1):42–58.
- Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10):551–560.
- Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.
- James Lee and Jason Lee. 2017. Shakespeare’s tragic social network; or why all the world’s a stage. *Digital Humanities Quarterly*, 11(2).
- John Lee and Chak Yan Yeung. 2012. Extracting networks of people and places from literary texts. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pages 209–218. Waseda University.
- Pádraig Mac Carron and Ralph Kenna. 2012. Universal properties of mythological networks. *Europhysics Letters*, 99(2):28002.
- Aibek Makazhanov, Denilson Barbosa, and Grzegorz Kondrak. 2014. Extracting family relationship networks from novels. *arXiv preprint arXiv:1405.0603*.
- Raymond A Mar and Keith Oatley. 2008. The function of fiction is the abstraction and simulation of social experience. *Perspectives on psychological science*, 3(3):173–192.
- Raymond A Mar, Keith Oatley, Jacob Hirsh, Jennifer Dela Paz, and Jordan B Peterson. 2006. Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of research in personality*, 40(5):694–712.
- Raymond A Mar, Keith Oatley, and Jordan B Peterson. 2009. Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes.
- Raymond A Mar, Jennifer L Tackett, and Chris Moore. 2010. Exposure to media and theory-of-mind development in preschoolers. *Cognitive Development*, 25(1):69–78.
- Philip Massey, Patrick Xia, David Bamman, and Noah A Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint arXiv:1512.00728*.
- Sandeep Soni David Bamman Andrew Piper Matthew Wilkens, Elizabeth F. Evans. forthcoming. Small worlds: Measuring the mobility of characters in english-language fiction. *Journal of Computational Literary Studies*, 3(1).
- Franco Moretti. 2011. Network theory, plot analysis. Technical report, Stanford Literary Lab.
- M Nijila and MT Kala. 2018. Extraction of relationship between characters in narrative summaries. In *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, pages 1–5. IEEE.
- Alan Palmer. 2004. *Fictional minds*. U of Nebraska Press.
- Andrew Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press.
- Andrew Piper. 2022. The conlit dataset of contemporary literature. *Open Humanities Data*, 8.
- Andrew Piper. 2024. What do characters do? the embodied agency of fictional characters. *Journal of Computational Literary Studies*, 2(1).
- Andrew Piper and Eva Portelance. 2016. How cultural capital works: Prizewinning novels, bestsellers, and the time of reading. *Post45*, 10.
- Andrew Piper, Hao Xu, and Eric D Kolaczyk. 2023. Modeling narrative revelation. In *CHR*, pages 500–511.
- Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas press.

Mia Ridge. 2016. *Making digital history: The impact of digitality on public participation and scholarly practices in historical research*. Open University (United Kingdom).

Henry Sauermann and Chiara Franzoni. 2015. Crowd science user contribution patterns and their implications. *Proceedings of the national academy of sciences*, 112(3):679–684.

Roel Smeets, Maarten De Pourcq, and Antal van den Bosch. 2021. Modeling conflict: Representations of social groups in present-day dutch literature. *Journal of Cultural Analytics*, 6:1–31.

Saatviga Sudhahar and Nello Cristianini. 2013. Automated analysis of narrative content for digital humanities. *International Journal of Advanced Computer Science*, 3(9):440–447.

Melissa Terras. 2015. Crowdsourcing in the digital humanities. *A new companion to digital humanities*, pages 420–438.

Marcello Trovati and James Brady. 2014. Towards an automated approach to extract and compare fictional networks: An initial evaluation. In *2014 25th International Workshop on Database and Expert Systems Applications*, pages 246–250. IEEE.

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. University of Chicago Press.

Beate Volker and Roel Smeets. 2020. Imagined social structures: Mirrors or alternatives? a comparison between networks of characters in contemporary dutch literature and networks of the population in the netherlands. *Poetics*, 79:101379.

Michaël C Waumans, Thibaut Nicodème, and Hugues Bersini. 2015. Topology analysis of social networks extracted from literature. *PloS one*, 10(6):e0126470.

Andrea Wiggins and Yurong He. 2016. Community-based data validation practices in citizen science. In *Proceedings of the 19th ACM Conference on computer-supported cooperative work & social computing*, pages 1548–1559.

Alex Woloch. 2009. *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton University Press.

Lisa Zunshine. 2006. *Why we read fiction: Theory of mind and the novel*. Ohio State University Press.

A Appendix: LLM Prompts

A.1 Base Prompt

What kind of interaction between char1 and char2? Choose one of six options: No, Associating, Thinking, Touching, Observing, Communicating.

A.2 Detailed Prompt

Task Description: Classify the type of interaction between char1 and char2 in a given passage. There are six categories of interaction:

No interaction: Direct or indirect interaction does not occur between char1 and char2. Any imagination or assumption of interaction also counts as No.

Communicating: char1 and char2 are engaged in some form of communication, such as speaking, writing, or signaling.

Associating: char1 and char2 are linked by a social or relational context, such as friendship, teamwork, or other associative bonds.

Observing: at least one character is observing or watching another one, without direct interaction.

Thinking: at least one character is thinking about or recalling memories of another one, without direct interaction.

Touching: char1 and char2 are engaged in physical touch or contact.

What kind of interaction between char1 and char2? Choose one of six options: No, Associating, Thinking, Touching, Observing, Communicating.

A.3 One-shot and Many-shot Prompt

Append examples with passage, char1, char2, and label before the detailed prompt. In the many-shot setting, any shot contains one example from each class.

Multi-word expressions in biomedical abstracts and their plain English adaptations

Sergei Bagdasarov

Saarland University, Germany
sergeiba@lst.uni-saarland.de

Elke Teich

Saarland University, Germany
e.teich@mx.uni-saarland.de

Abstract

This study analyzes the use of multi-word expressions (MWEs), prefabricated sequences of words (e.g. *in this case, this means that, health-care service, follow up*) in biomedical abstracts and their plain language adaptations. While English academic writing became highly specialized and complex from the late 19th century onwards, recent decades have seen a rising demand for a lay-friendly language in scientific content, especially in the health domain, to bridge a communication gap between experts and laypersons. Based on previous research showing that MWEs are easier to process than non-formulaic word sequences of comparable length, we hypothesize that they can potentially be used to create a more reader-friendly language. Our preliminary results suggest some significant differences between complex and plain abstracts when it comes to the usage patterns and informational load of MWEs.

1 Introduction

Previous diachronic research has shown that English scientific writing developed a compressed code of communication that is efficient for its primary users (i.e. scientists) (Halliday, 1988; Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2016, 2022). However, the consequence of this process was that academic papers became almost incomprehensible for a general audience, which poses a considerable problem as the need to draw knowledge directly from scientific publications is growing among laypersons, particularly in the health domain.

To address this issue, many scholars and journals encourage scientists to use plain language in their papers or at least include plain language summaries of their work (Hauck, 2019; Sedgwick et al., 2021). While writing recommendations on plain language abound, they seem to pay little attention to multi-word expressions (MWEs), i.e. prefabricated sequences of several words that are argued to

foster the fluency of language use (Sinclair, 1991; Pawley and Syder, 1983).

Our goal is to investigate whether, and if so to what extent MWEs can ease the processing of plain language texts. To this end, we analyze the use of MWEs in abstracts from biomedical papers ("complex abstracts") and their plain language adaptations ("plain abstracts"). We pose the following questions: (i) Does the use of MWEs differ in complex and plain abstracts and, more specifically, do plain abstracts use more MWEs? (ii) Are MWEs in plain abstracts easier to process?

In general, we expect to see more MWEs in plain abstracts. In terms of MWE types, we anticipate that nominal MWEs (e.g. compound nouns used as terms), typically associated with technical scientific writing, will be less characteristic of plain abstracts. Moreover, we expect MWEs in plain abstracts to be less informationally loaded on average (and therefore easier to process).

The remainder of the paper is structured as follows. Section 2 is dedicated to MWE processing. Section 3 describes our data and methodology. In Section 4, we present our analysis results. In Section 5, we provide a summary and prospects of future work.

2 Background and Related Work

Linguistic studies in recent decades have revealed that MWEs make up a large proportion of language use and that they are less costly in processing than other sequences of words (Erman and Warren, 2000; Foster, 2001). For instance, Conklin and Schmitt (2008) prove that MWEs have shorter reading times in comparison to non-formulaic expressions. Li et al. (2021) and Siyanova-Chanturia et al. (2011) arrive at similar conclusions using eye-tracking. The assumption about a processing advantage of MWEs has also been corroborated by EEG studies (cf. Tremblay et al. (2011); Siyanova-Chanturia et al. (2017)). Further evidence is pro-

vided from speech processing: formulaic expressions are produced faster and more fluently than comparable, non-formulaic expressions and recognized better (e.g. under acoustic degradation; (Rammell et al., 2017)).

While it now seems increasingly clear that MWEs are faster and easier to process than non-formulaic language, what still remains open is whether the use of MWEs is influenced by other factors. In register theory, it is widely assumed that speakers adjust their language according to the particular communicative situation (Biber, 2012; Biber and Conrad, 2019; Conrad and Biber, 2005). One of the parameters describing the communicative situation is the relationship between the speaker and the recipient. For instance, in case of complex abstracts both speaker and recipient have professional knowledge of the subject. In contrast, plain abstracts are written by well-versed speakers for lay recipients. Hence, it is plausible to suppose that this shift in the level of expertise should be reflected in the use of MWEs, i.e. MWEs should be employed in plain abstracts in such a way that reduces the processing cost for the recipient.

3 Methodology

3.1 Data

We use the Plain Language Adaptation of Biomedical Abstracts dataset (PLABA) (Attal et al., 2023). The biomedical abstracts come from PubMed and were transformed into plain language by human writers on a sentence basis, with sometimes multiple plain language adaptations being written for one complex abstract. Some relevant corpus statistics is summarized in Table 1.¹

	#Abstracts	#Tokens	#Types
Complex	749	199,851	17,425
Plain	919	249,301	13,117

Table 1: PLABA corpus data

We performed tokenization and sentence segmentation with TreeTagger (Schmid, 1994, 1995). The pretokenized abstracts were then parsed with the state-of-the-art Stanza parser (Qi et al., 2020).

¹The number of abstracts available in PLABA at the time of our study differs from the number of abstracts stated in the original publication by Attal et al. (2023). Table 1 contains statistics on the actual data employed in our study.

3.2 MWE Identification

Following Alves et al. (2024a,b), we use Universal Dependencies (UD) and the Academic Formulas List (AFL) to identify MWEs in our corpus.

The UD framework (de Marneffe et al., 2021) contains five MWE-related labels: 1. compound — combinations of tokens that morphosyntactically behave as single words; in English this label refers mostly to nominal compounds (e.g. *muscle cramps*), 2. compound:prt — phrasal verbs (e.g. *follow up*), 3. fixed — certain grammaticized expressions normally acting as function words (e.g. *according to*), 4. flat — sequences where none of the words can be identified as the head, in our case these are mostly proper names (e.g. *Moderna mRNA-1273*), 5. flat:foreign — sequences of foreign words².

The identification of MWEs according to the UD method was performed using a Python script that extracted all words labelled with the above mentioned tags and their corresponding heads (if any). For instance, some occurrences of the word *muscle* were labelled with the *compound* tag during parsing, with the word *cramps* being identified as their head. So, the resulting MWE is *muscle cramps*.

The AFL (Simpson-Vlach and Ellis, 2010) includes 207 core formulaic expressions common for both written and spoken academic English, 200 expressions common for written academic English and 200 expressions common for spoken academic English. The authors selected the MWEs based on a measure called "formula teaching worth", which combines frequency and mutual information. For this study, we relied only on core and written MWEs. Using a Python script, we iterated through both lists and extracted all MWEs that appear at least once in our data.

After applying the UD and AFL methods, we merged all extracted MWEs into one final list. No frequency thresholds were used since the UD labels are grammatically motivated and the AFL MWEs had already been predefined based on specific measures.

3.3 Relative Entropy

We use the asymmetric variant of relative entropy, known as Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951), to investigate the use of MWEs in complex and plain abstracts. KLD allows us to compare two probability distributions

²This category is not attested in our data.

A and B (here, MWEs in complex and plain abstracts) by showing the number of additional bits of information needed to encode one distribution using the other one. The formal representation of KLD is shown in equation 1:

$$D(A||B) = \sum_i p(feature_i|A) \log_2 \frac{p(feature_i|A)}{p(feature_i|B)} \quad (1)$$

A KLD value of 0 would mean that the usage patterns of MWEs are exactly the same in complex and plain abstracts, while a value greater than 0 would indicate a divergence. Moreover, KLD shows the contributions of individual features to the overall divergence, allowing us to generate a list of the most relevant features (i.e. MWEs).

3.4 Surprisal

To quantify the informativity of MWEs, we use surprisal, a measure that shows how much information (in bits) a word carries in a given context (Shannon, 1948):

$$S(word) = -\log_2 p(word|context) \quad (2)$$

As shown by reading time or specific EEG signals, surprisal is proportional to cognitive effort. Hence, a high surprisal of a MWE would be indicative of its high processing cost and vice versa.

In this study, we estimated surprisal of a given word n based on the four-gram model where words $n-1$, $n-2$ and $n-3$ are taken as context (cf. Genzel and Charniak (2002)). Additionally, we computed average surprisal for each MWE. For this, we first estimated average surprisal of each individual instance of a MWE, then summed all values and divided them by the number of occurrences of a MWE in the corpus.

4 Results

Contrary to what we expected, plain abstracts employ fewer MWEs, both in terms of unique occurrences (6,155 vs 6,700) and total frequency (62,976.08 vs 63,802.53 occurrences per million words). Compounds are the most common MWE type in both abstract categories as reflected in Table 2 and Figure 1. The most notorious differences in frequencies were observed for the proper nouns (flat) and phrasal verbs (compound:prt).

However, going beyond mere frequency estimations, our KLD analysis revealed a considerable difference in the use of MWEs in both directions

Type	Description	Complex	Plain
compound	compounds	6,309	5,702
compound:prt	phrasal verbs	26	114
fixed	fixed expressions	30	35
flat	proper names	67	29
afl	academic formulas	268	275

Table 2: Unique MWEs identified in complex and plain abstracts

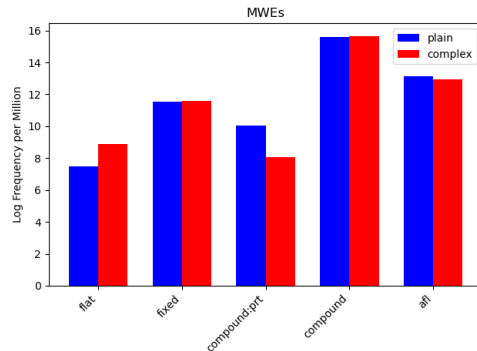


Figure 1: MWE frequency in complex and plain abstracts.

of comparison: 3.35 bits and 3.18 bits for complex VS plain and plain VS complex, respectively. A look at the most distinctive features (see Figures 2 and 3) also offers interesting insights. While both types of abstracts are characterized by compound MWEs to a great extent, we see, for instance, that complex abstracts have more statistical terminology (e.g. *confidence interval*, *mean age*, *odds ratio* etc.) and different research design terms (e.g. *cohort study*, *crossover study*, *control group* etc.).

In contrast, such MWEs are not encountered among the features distinctive of plain abstracts. This is probably due to the fact that such statistical and methodological information is not relevant for a lay person and, therefore, can be left out to enhance readability.

Moreover, we see numerous examples of specialized terminology denoting biological and medical phenomena (e.g. *dopamine receptor*, *plasma concentration* etc.). Since it is impossible to just delete such terms without losing information relevant to the reader, plain abstracts try to use more common equivalents (e.g. *blood sugar levels* instead of *blood glucose levels*). Sometimes such transformations lead to the creation of MWEs in cases where no MWE is used in complex abstracts. For instance, a one-word term *placebo* turns into a compound *dummy treatment*. Or an adjective-noun term *neurodegenerative disease* is replaced with a



Figure 2: 70 most distinctive MWEs in complex abstracts.

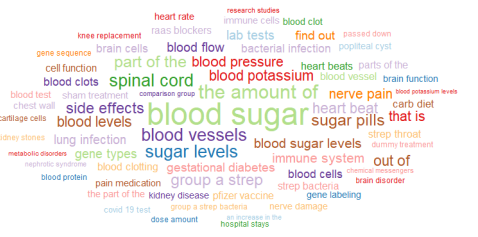


Figure 3: 70 most distinctive MWEs in plain abstracts.

noun compound *brain disorder*.

Other MWE types are also attested among the most characteristic features, albeit marginally. For instance, phrasal verbs seem to be more distinctive of plain abstracts (e.g. *find out, make up*). Fixed and AFL MWEs are present in both lists, however complex abstracts seem to employ more sophisticated expressions typical of elaborated writing (e.g. *as well as, in terms of, according to the* etc.). Flat MWEs are not attested among the most distinctive features.

In terms of informativity, as measured by surprisal, we observed significant differences between complex and plain abstracts for compound and AFL categories, while phrasal verbs showed a marginally significant difference.³ All of these three MWE types have lower surprisal for plain abstracts (see Figure 4), which is, in principle, in line with our expectations although we anticipated a more pronounced trend.

Lower surprisal values in plain compound MWEs might be indicative of MWEs being used to effectively reduce processing effort. Consider, for instance, the MWE *blood glucose levels* which is typically used in complex abstracts and has a surprisal of 4.73 bits. Its plain language equivalent *blood sugar levels*, however, transmits 3.24 bits of information on average, thus being easier to process.

The same seems to hold for cases where plain abstracts use an MWE instead of a noun with an adjective premodifier, which is a very common terminology formation pattern in scientific texts. Compare, for example, the following terms from complex abstracts and their plain language equiv-

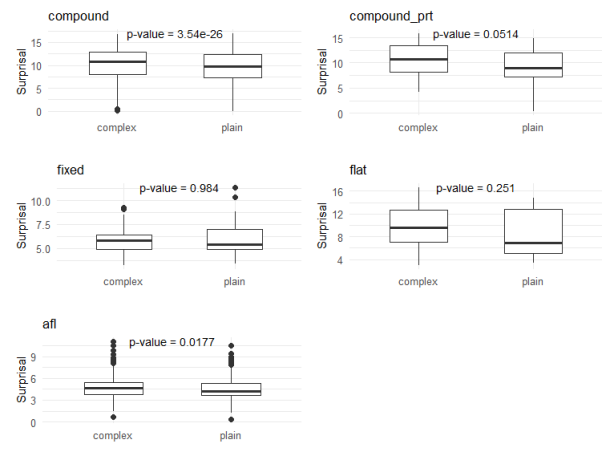


Figure 4: Comparison of average MWE surprisal in complex and plain abstracts across different MWE types.

alents: *renal cancer* (7.41 bits) vs *kidney cancer* (5.91 bits) and *neurodegenerative disease* (12.20 bits) vs *brain disorder* (3.65 bits).

A similar mechanism seems to apply to phrasal verb MWEs that can be used as an alternative to more complex verbs. Consider, for instance, Example (1) extracted from a complex abstract and its plain language adaptation shown in Example (2) (values in parenthesis indicate surprisal in the corresponding sentence)

- (1) ... *as the risk of detrimental outcomes **increases*** (12.46) *with delayed surgical intervention.*
- (2) ...*since the risk of harmful effects **goes up*** (8.69) *with delayed surgery.*

While in general our findings do suggest that some types of MWEs *per se* seem to be easier to process in plain language abstracts, a more in-depth analysis is needed to investigate how rewrit-

³For determining statistical significance we used the Wilcoxon rank sum test available in R (R Development Core Team).

ing strategies like those described above affect processing complexity on the sentence and text level.

5 Conclusion and Future Work

In this study, we investigated MWEs in biomedical abstracts and their plain language adaptations. We were able to establish some differences in the use of MWEs (e.g. more prominent use of statistical and methodological terms in complex abstracts, greater reliance on phrasal verbs in plain abstracts).

Furthermore, we found that the informational load of compound, phrasal verb and AFL MWEs is lower in plain abstracts, suggesting that the use of MWEs might play a role in decreasing processing cost in the transition from complex to plain language.

In future studies, we will focus on the MWE types individually to investigate why plain MWEs have lower surprisal. Additionally, we are planning to expand our methodology to account for factors that might be correlated with the MWE processing cost (e.g. association strength among the component parts of an MWE).

Limitations

Our study is based on a relatively small corpus: roughly 200,000 words for complex abstracts and 250,000 words for plain abstracts. Moreover, we are not aware which journals the abstracts come from and whether the authors of abstracts are native speakers of English. These factors might also influence the use of MWEs. The creation of a larger dataset with detailed meta-information may be addressed in future research.

Ethics Statement

This does not apply to our research since we did not perform any experiments nor collected personal data.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt, and Elke Teich. 2024a. [Diachronic analysis of multi-word expression functional categories in scientific English](#). In *Proceedings of the Joint*

Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, pages 81–87, Torino, Italia. ELRA and ICCL.

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024b. [Multi-word expressions in English scientific writing](#). In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 67–76, St. Julians, Malta. Association for Computational Linguistics.

Kelly Attal, Brian Ondov, and Dina Demner-Fushman. 2023. [A dataset for plain language adaptation of biomedical abstracts](#). *Scientific Data*, 10(1):8.

Douglas Biber. 2012. [Register as a predictor of linguistic variation](#). *Corpus Linguistics and Linguistic Theory*, 8(1):9–37.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2nd edition. Cambridge University Press, Cambridge.

Douglas Biber and Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Studies in English Language. Cambridge University Press, Cambridge, UK.

Kathy Conklin and Norbert Schmitt. 2008. [Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?](#) *Applied Linguistics*, 29(1):72–89.

Susan Conrad and Douglas Biber. 2005. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, 20(2004):56–71.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.

Stefania Degaetano-Ortlieb and Elke Teich. 2016. [Information-based modeling of diachronic linguistic change: From typicality to productivity](#). In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 165–173, Berlin, Germany. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. [Toward an optimal code for communication: The case of scientific English](#). *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

Britt Erman and Beatrice Warren. 2000. [The idiom principle and the open choice principle](#). *Text & Talk*, 20(1):29–62.

Pauline Foster. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In Martin

- Bygate, Peter Skehan, and Merrill Swain, editors, *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*, pages 75–93. Longman, Harlow.
- Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th ACL*, pages 199–206, Philadelphia, PA, USA.
- M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of written English: Situational factors and linguistic features*, pages 162–177. Pinter, London.
- Steven A. Hauck. 2019. [Sharing planetary science in plain language](#). *Journal of Geophysical Research: Planets*, 124(10):2462–2464.
- Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Hui Li, Kate L. Warrington, Ascensión Pagán, Kevin B. Paterson, and Xinchun Wang. 2021. [Independent effects of collocation strength and contextual predictability on eye movements in reading](#). *Language, Cognition and Neuroscience*, 36(8):1001–1009.
- Andrew Pawley and Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards and Richard W. Schmidt, editors, *Language and communication*, pages 191–225. Longman, London, UK.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python Natural Language Processing Toolkit for Many Human Languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- R Development Core Team. [R: A language and environment for statistical computing](#). R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- C. Sophie Rammell, Diana Van Lancker Sidsis, and David B. Pisoni. 2017. Perception of formulaic and novel expressions under acoustic degradation. *Ment Lex*, 12(2):234–262.
- Helmut Schmid. 1994. [Probabilistic part-of-speech tagging using decision trees](#). In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.
- Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Kyoto, Japan.
- Cassie Sedgwick, Laura Belmonte, Amanda Margolis, Patricia Osborn Shafer, Jennifer Pitterle, and Barry E. Gidal. 2021. [Extending the reach of science – talk in plain language](#). *Epilepsy Behavior Reports*, 16:100493.
- Claude Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27:379–423, 623–656.
- Rita Simpson-Vlach and Nick C. Ellis. 2010. [An Academic Formulas List: New Methods in Phraseology Research](#). *Applied Linguistics*, 31(4):487–512.
- John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.
- Anna Siyanova-Chanturia, Kathy Conklin, Sendy Caffarra, Edith Kaan, and Walter J.B. van Heuven. 2017. [Representation and processing of multi-word expressions in the brain](#). *Brain and Language*, 175:111–122.
- Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. 2011. [Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers](#). *Second Language Research*, 27(2):251–272.
- Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language learning*, 61(2):569–613.

Assessing the Performance of ChatGPT-4, Fine-tuned BERT and Traditional ML Models on Moroccan Arabic Sentiment Analysis

Mohamed Hannani
University of Siegen, Germany
mohamed_hannani@yahoo.com

Abdelhadi Souidi
Ecole Nationale Supérieure
des Mines de
Rabat, Morocco
asouidi@enim.ac.ma

Kristof Van Laerhoven
University of Siegen, Germany
kvl@eti.uni-siegen.de

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in various natural language processing tasks across different languages. However, their performance in low-resource languages and dialects, such as Moroccan Arabic (MA), requires further investigation. This study evaluates the performance of ChatGPT-4, different fine-tuned BERT models, FastText as text representation, and traditional machine learning models on MA sentiment analysis. Experiments were done on two open source MA datasets: an X(Twitter) Moroccan Arabic corpus (MAC) and a Moroccan Arabic YouTube corpus (MYC) datasets to assess their capabilities on sentiment text classification. We compare the performance of fully fine-tuned and pre-trained Arabic BERT-based models with ChatGPT-4 in zero-shot settings.

Keywords: Sentiment Analysis, Bert, GPT, Moroccan Arabic, LangChain

1 Introduction

The field of sentiment analysis (SA) has made remarkable advances, enabling the extraction and analysis of human sentiments from textual data for a variety of purposes. This technology has diverse applications, ranging from social media monitoring and market research to political discourse analysis.

However, SA faces several challenges, such as the phenomena of data and concept drift (Zhao et al., 2022), particularly pronounced in the ever-evolving landscape of social media. Data drift refers to changes in the statistical properties of the input data over time, while concept drift indicates a deeper shift in the underlying meaning or interpretation of the data that the model aims to predict. These drifts can lead to a decrease in the accuracy of SA models if not addressed. For instance, the way people express sentiments online can evolve rapidly, making previously trained models less effective. This necessitates ongoing monitoring and

adaptation of SA methodologies, including updating rule-based systems and dictionaries, as well as retraining machine learning models to ensure they remain aligned with the shifting linguistic and cultural contexts of web-based communication.

While SA has seen substantial progress in major languages, its application to dialectal languages, such as MA, a regional variant of Modern Standard Arabic (MSA), has not received much attention. MA is the main medium of communication among Moroccans. The unique linguistic features of MA, including regional variations, colloquialism, borrowed words from other languages, coupled with the use of multiple scripts (Arabic and Latin), present significant challenges for SA. Previous SA research (Elmadany et al., 2022) has predominantly focused on corpora written in Arabic script despite the increasing prevalence of Latin script usage in online communication, particularly on social media platforms.

Recent advancements in AI, particularly the emergence of LLMs, such as GPT-4 (OpenAI, 2023), PaLM 2 (Anil et al., 2023) and Falcon (Penedo et al., 2023) offer potential solutions to the challenges posed by data and concept drift in SA. These models, trained on vast and diverse datasets and fine-tuned for various tasks, have demonstrated promising capabilities in SA (Wang et al., 2023; Inoue et al., 2021; Amin et al., 2023). While some research has explored the potential of LLMs for Standard Arabic sentiment analysis (Al-Thubaity et al., 2023), no study has so far evaluated the performance of LLMs on MA. This work conducts the first-ever evaluation of ChatGPT's performance on MA SA, offering valuable insights into the applicability of LLMs in analyzing sentiment in Arabic dialects.

2 Related Work

The availability of sentiment data from social media platforms has greatly increased interest in Arabic sentiment analysis (SA) research over the last ten years. Speakers of Arabic dialects were historically limited to using their dialects only when speaking. However, the emergence of social media has given Arabic speakers the ability and space to express themselves in writing as well (Darwish et al., 2021). This has resulted in an abundance of informal, dialectal textual material, as opposed to MSA formality. A multitude of datasets spanning multiple genres—mostly tweets—have been created for Arabic SA, including Egyptian (Nabil et al., 2015; Refaee and Rieser, 2014), Levantine (Baly et al., 2019), Maghrebi (Mdhaftar et al., 2017), as well as the Saudi dialect (Assiri et al., 2016). Other datasets (Al-Obaidi and Samawi, 2016; Abdul-Mageed et al., 2014) include several Arabic dialects in addition to MSA.

Arabic Sentiment Analysis has traditionally concentrated on rule-based techniques, much like other languages (ElSahar and El-Beltagy, 2014; Al-Twairish et al., 2016). The main goal of these techniques was to create sentiment lexicons. Arabic Sentiment Analysis has seen a rise in interest in applying machine learning techniques in recent years. These techniques are less vulnerable to the drawbacks of lexicon-based techniques and are capable of identifying sentiment patterns from a big corpus of text. To implement morphological and syntactic features, popular machine learning techniques have been employed, such as Naïve Bayes (NB), Support Vector Machines (SVMs), and K-Nearest Neighbor (kNN) classifiers (Abdul-Mageed et al., 2014; Duwairi and Qarqaz, 2014; Abdulla et al., 2013).

Transformer-based models, such as BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019), have been introduced and proved successful in several natural language processing (NLP) applications. BERT and BERT-like models achieved state-of-the-art performance on many NLP tasks, including sentiment analysis in many languages (Sun et al., 2019).

Abdul-Mageed et al. (2020); Antoun et al. developed two models, ARBERT and MARBERT, pre-trained on a large collection of datasets in MSA and several Arabic dialects (Levantine, Moroccan Arabic, etc.). They reported new state-of-the-art results

on the majority of the datasets in their fine-tuning benchmark.

In addition to discriminative models, such as BERT, generative models have recently gained prominence in NLP research. These models, such as GPT (Radford et al., 2018; Brown et al., 2020), T5 (Raffel et al., 2020), and BLOOM (Scao et al., 2022), are designed to create new text samples. Multilingual and language-specific versions of these models have been developed. For example, AraT5 (Elmadany et al., 2022) and AraGPT-2 (Antoun et al.) are tailored for Arabic. Generative models have demonstrated potential in tasks, such as text completion, translation, summarization, and even sentiment analysis, where they can generate text that aligns with specific sentiments (Al-Thubaity et al., 2023).

In this work, we evaluate the performance of ChatGPT-4 and transformer-based models on SA of MA using the aforementioned open source datasets, namely MAC (Garouani and Kharroubi, 2021) and MYC (Jbel et al., 2023). To our knowledge, this is the first-ever attempt to compare the performance of these models on the MA SA task.

The structure of the rest of the paper is as follows: Section 3 presents the experimental setup and the various experiments we conducted, datasets, and model architectures. Section 4 presents the experiments' results and analysis. Section 5 provides concluding remarks and future work.

3 Experimental Setup

Our research objective is to evaluate the capability of the ChatGPT model, some existing pre-trained BERT models, and FastText (Joulin et al., 2017) as sentiment analyzers for MA. To assess this, we utilize the two datasets, MAC and MYC, designed for sentiment analysis. On each dataset, we evaluate ChatGPT, fine-tuned BERT-based, and FastText models. Furthermore, we compared the fine-tuned and pre-trained BERT-based models with the ChatGPT results. The results of our experiments are compared with other related work.

We use gpt-4-turbo model by OpenAI¹ for both MAC and MYC datasets as a sentiment analyzer for MA. We asked the model to predict the class of the given input tokens. Table 1 summarizes the parameters and the prompt used when calling the model.

The primary objective of our experiments is to

¹<https://openai.com/>

Parameters	temperature	top_p
Values	0.3	1

Table 1: GPT-4-turbo parameters with OpenAI API

assess the capabilities of generative models and BERT-based models, as well as FastText for MA sentiment analysis. We evaluate the following models:

- GPT-4, accessed via ChatGPT by OpenAI,
- Pre-trained/fine-tuned BERT-based models,
- FastText as text representation.
- Traditional machine learning classifiers

For GPT-4, we utilize the ChatOpenAI wrapper provided by LangChain framework ² to send prompts and receive responses. For BERT-based models, we fine-tune (full network or freezing the model’s backbone) various existing models pre-trained on a large corpus in a variety of languages. Table 2 provides information on the BERT-based models we used for our experiments.

3.1 Prompt Composition

The system prompts used for calling the GPT model for the MAC and MYC datasets are presented below.

MAC Dataset Prompt

```
<Predict the class of this Arabic review (e.g ternary classification), whether it's positive (return 2), neutral (return 1) or negative (return 0) review. Please do not return anything other than that.>
```

MYC Dataset Prompt

```
<Predict the class of this Arabic review (e.g binary classification), whether it's positive (return 2) or negative (return 0) review. Please do not return anything other than that.>
```

To facilitate prompt composition and enhance sentiment detection, we integrated LangChain into our system. LangChain serves as a framework designed for the development of applications leveraging LLMs. Its primary objective is to empower

²<https://www.langchain.com/>

developers with the seamless integration of diverse data sources and the facilitation of interactions with other applications. To achieve this goal, LangChain framework offers modular components, serving as abstractions, and customizable use of case-specific pipelines, referred to as chains. We also used a json parser as part of the Chain-of-Thoughts to ensure getting exactly and only the class label when invoking the model API. The prompt template is shown in Figure 1.

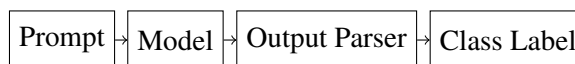


Figure 1: Chain-of-Thoughts used for Sentiment Analysis in Moroccan Arabic Dialect

The Prompt is the SystemMessage component followed by HumanMessage from Langchain framework. The Prompt is then used to request the API, and the API response is then sent to the JsonOutputToolsParser provided by the same framework to parse the response for consistency.

3.2 Sentiment Datasets

For the aforementioned experiments, we use two datasets: the MAC dataset, an MA corpus consisting of 18000 manually labeled tweets, resulting in a lexicon-dictionary of 30000 words labeled as positive, negative, and neutral. Table 3 below shows information about the pre-processed MAC dataset. We had to remove tweets that have the class mixed (MSA and MA) and missing values of the type column (when not labeled).

The complexity of Moroccan web content features a blend of Arabic and Latin script. This dual-script usage in MA adds a layer of complexity that traditional sentiment analysis approaches might overlook. To evaluate our models on that type of MA, we used the MYC dataset (Jbel et al., 2023), which contains 20k (raw data) comments scrapped from 50 Moroccan famous YouTube channels on different topics. Table 3 below showcases statistics about the pre-processed MYC dataset.

It is worth mentioning that the pre-processed MYC shared by Jbel et al. (2023) is not really pre-processed as they claimed in their paper. We tried to follow their pre-processing steps in their paper, namely, remove empty comments, remove usernames, remove links (https and http links), and remove unlabeled samples.

Model Name	Pre-training Language	Vocabulary Size
bert-base-multilingual-cased	Multilingual	119547
bert-base-arabic	Arabic	32000
darijabert-arabizi	Arabic	110000
DarijaBERT	Moroccan Arabic Dialect (Darija)	80000
bert-base-arabertv2	Arabic	64000

Table 2: BERT-Based Models and Pre-training Languages

Dataset	Size	Tweet Class			Arabic Type	
		Positive	Negative	Neutral	Modern Standard Arabic	Moroccan Arabic
MAC	18k	9888	3505	4039	12145	5287
MYC	16k	7427	8621	-	-	-

Table 3: The pre-processed MAC and MYC Dataset statistics

3.3 Models Architecture And Setup

For our experiments with BERT-based and FastText models, we employed a custom classifier head, as illustrated in Figure 2 which consists of a sequential architecture incorporating linear transformations, ReLU activation, Dropout for regularization, and a final Softmax layer for classification. This classifier head was integrated with powerful pre-trained language models like BERT, known for its contextual understanding capabilities. Specifically, the BERT model’s output from the pooler was fed into our classifier, allowing us to leverage BERT’s deep semantic representations. Additionally, we incorporated fastText embeddings, renowned for their efficiency in handling morphologically rich languages like Arabic.

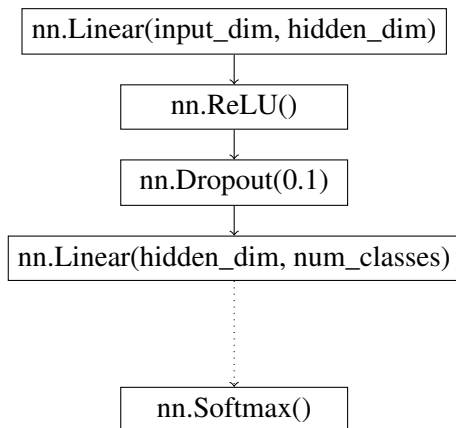


Figure 2: Classifier head used for BERT-based models and FastText for text representation for SA.

For the BERT-based models shown in Table 2, we used the BERT backbone (frozen or trained from scratch), the pooler output is fed to the classifier shown in Figure 2, and the Table 4 shows the

parameter values of the classifier head when used with BERT-based models.

input_dim	hidden_dim	num_classes
768	512	3(MAC), 2(MYC)

Table 4: classifier head’s parameters with BERT-based models

For the FastText model, we used facebook/fasttext-ar-vectors (Joulin et al., 2016). FastText is a library for efficient learning of word representations and sentence classification. Table 5 shows the parameter values of the classifier head when used with FastText model.

input_dim	hidden_dim	num_classes
300	128	3(MAC), 2(MYC)

Table 5: classifier head’s parameters with FastText model

For the ChatGPT model, we evaluated the GPT-4-Turbo model on the pre-processed \mathbf{MAC}_{full} and \mathbf{MYC}_{full} with the parameters shown previously in Table 1 and prompt as discussed in Section 3.1 for both datasets to see their performance on different datasets with different types and sources (Twitter and YouTube).

For BERT-based models, we trained all the mentioned pre-trained BERT models shown in Table 2, We used stratified sampling to ensure a balanced class distribution across test and train sets. We used $\mathbf{MAC}_{80\%}$ for training and $\mathbf{MAC}_{20\%}$ for evaluation. The same approach was applied to the MYC dataset (ie., $\mathbf{MYC}_{80\%}$ for training and $\mathbf{MYC}_{20\%}$ for testing).

For the FastText model, we used $\text{MAC}_{80\%}$, $\text{MYC}_{80\%}$ for training, and $\text{MAC}_{20\%}$, $\text{MYC}_{20\%}$ for testing.

In the following section 4, we present the results of the experiments, compare, and discuss the evaluation results for each model and dataset.

4 Results And Analysis

4.1 GPT-4 model via OpenAI

As discussed previously in Section 3.1 and 3.3, we evaluated the gpt-4-turbo model on MAC_{full} and MYC_{full} datasets, as well as on test subsets, $\text{MAC}_{20\%}$ and $\text{MYC}_{20\%}$ for comparison purposes with other models.

Since the MAC dataset contains tweets in MSA and MA, we also aimed to evaluate the performance of the model in each class. The following Table 6 summarizes the results of the evaluation in MAC_{full} and MAC_{test} with 20% of the dataset.

Accuracy	Precision	Recall	F1-score
MAC_{full}			
Modern Standard Arabic			
0.710	0.741	0.710	0.713
Moroccan Arabic			
0.690	0.714	0.690	0.688
MAC_{20%}			
Modern Standard Arabic			
0.635	0.720	0.635	0.654
Moroccan Arabic			
0.607	0.673	0.607	0.620

Table 6: GPT-4 model performance in MAC_{full} and $\text{MAC}_{20\%}$ across the type of class (MSA or MA).

As can be seen in Table 6, for both full and test sets, we notice that the GPT-4 model performs well in tweets written in MSA, compared to those written in (MA). This difference in performance can be attributed to several factors. Firstly, language uniformity plays a significant role. MSA is a standardized and formal version of Arabic used in official communication, media, literature, and formal speeches. It has consistent grammar, vocabulary, and syntax, which makes it easier for NLP models to learn and predict accurately. In contrast, MA varies significantly across regions and often incorporates local slang, colloquialisms, and foreign words. This linguistic diversity and lack of standardization make it challenging for models to perform consistently. Secondly, the availability

and quality of training data influence model performance. Models, such as GPT-4 are often trained on large corpora that include a substantial amount of MSA texts, given its prevalence in written and formal contexts. This extensive training on MSA helps the model learn its patterns more effectively. On the other hand, there is generally less training data available for dialectical variants due to their informal use and the vast regional differences. This scarcity of training data can lead to poorer model performance on dialectical texts.

In the $\text{MAC}_{20\%}$ subset, the metrics for the Standard Arabic classifier show an accuracy of 0.6356, precision of 0.7205, recall of 0.6356, and F1 score of 0.6545.

To understand the challenges the model faces in classifying tweets, we examined the performance across different sentiment classes: positive, neutral, and negative. Table 7 shows the scores across each class (positive, negative, and neutral) on $\text{MAC}_{20\%}$ set.

Class	Precision	Recall	F1 Score
Modern Standard Arabic			
Negative	0.80	0.85	0.82
Neutral	0.35	0.61	0.44
Positive	0.84	0.57	0.68
Moroccan Arabic			
Negative	0.55	0.81	0.66
Neutral	0.38	0.50	0.43
Positive	0.85	0.58	0.69

Table 7: Performance metrics for Modern Standard Arabic and Moroccan Arabic tweets on $\text{MAC}_{20\%}$ set.

The classification reports 7 offer detailed insights into the performance of the model, providing metrics such as precision, recall, and F1-score for each sentiment class. With respect to MSA, the model demonstrates strong performance in identifying negative tweets, achieving high precision (0.80) and recall (0.85). However, it struggles with neutral tweets, as evidenced by the lower precision (0.35) and recall (0.61), indicating difficulty in distinguishing neutral sentiment. Similarly, while the model exhibits high precision (0.84) in classifying positive tweets, the lower recall (0.57) suggests that some positive tweets are misclassified as neutral or negative.

In the case of MA, the model achieves moderate precision (0.55) and recall (0.81) for negative tweets, indicating reasonable performance in this

class. However, the precision (0.38) and recall (0.50) for neutral tweets are significantly lower, highlighting challenges in accurately predicting neutral sentiment. Despite maintaining high precision (0.85) for positive tweets, similar to MSA, the model struggles with recall (0.58), indicating misclassification issues.

A key observation from these reports is the consistent difficulty the model encounters with neutral tweets across both MSA and MA. Lower precision and recall scores suggest that neutral tweets are often misclassified as either positive or negative, indicating a need for improved classification strategies for neutral sentiment. Additionally, while the model generally performs better on negative and positive classes, the lower recall for positive tweets suggests a tendency to miss some positive instances, possibly predicting them as neutral or negative.

In the remainder of this section, we evaluate the performance of GPT-4 on the MYC dataset, purely dialectal data. Unlike the MAC dataset, MYC includes both Arabic and Latin script, as discussed previously in Section 3.2

The following Table 8 summarizes the results of the evaluation on the full MYC dataset.

Accuracy	Precision	Recall	F1-score
MYC_{full}			
0.624	0.623	0.622	0.622
MYC_{20%}			
0.608	0.6087	0.608	0.607

Table 8: GPT-4 model performance in **MYC_{full}** and **MYC_{20%}**

To understand the challenges encountered by the model in categorizing tweets, it is essential to analyze its performance across various sentiment categories: positive and negative. Table 9 presents the performance metrics for each sentiment class (positive and negative) on the **MYC_{20%}** dataset. The performance metrics for sentiment classification on the **MYC_{20%}** dataset, as shown in Table 9, highlight varying degrees of success in accurately classifying tweets into positive and negative sentiment categories. The model achieves a precision of 0.647 and a recall of 0.655 for negative tweets, indicating a relatively balanced ability to correctly identify negative sentiment instances while minimizing false negatives. Conversely, for positive tweets, the precision is 0.597, indicating that a sig-

nificant portion of the positively classified tweets may be incorrect, while the recall is 0.588, suggesting a lower ability to capture all positive instances present in the **MYC_{full}** dataset.

Class	Precision	Recall	F1 Score
Negative	0.647	0.655	0.651
Positive	0.597	0.588	0.593

Table 9: Performance on **MYC_{full}** across each class.

4.2 Fine-tuned BERT-based models

In this section, we explore the performance of various BERT-based models presented previously in Table 2 trained and evaluated on MAC and MYC datasets and configurations. More specifically, we conducted experiments with the following configurations:

1. BERT-based models trained on **MAC_{80%}** and evaluated on **MAC_{20%}** dataset. This model was trained with two options: (a) training the entire network, and (b) freezing the backbone and training only the classifier.
2. BERT-based models were fully trained on **MAC_{80%}** and evaluated on **MYC_{20%}** dataset.
3. BERT-based models were fully trained on **MYC_{80%}** and evaluated on **MYC_{20%}** dataset.

The performance metrics of the BERT-based models trained and evaluated on the MAC dataset, as shown in Table 10, highlight significant differences between models trained with fully unfrozen networks and those with frozen backbones. When the entire network is trained, models such as DarijaBERT and bert-base-arabertv2 demonstrate superior performance, with DarijaBERT achieving the highest accuracy of 0.90, precision of 0.881, and F1-score of 0.877. This indicates a robust capability to capture the nuances of the MAC dataset. Conversely, models trained with frozen backbones exhibit notably lower performance, with the bert-base-multilingual-cased model showing the lowest accuracy (0.602) and F1-score (0.353).

To further evaluate the generalization capabilities of our BERT-based models, we conducted experiments where the models were fully trained on **MAC_{80%}** and evaluated on **MYC_{20%}** dataset. This approach allows us to assess how well the models, trained on Twitter data (MAC), perform when

Accuracy	Precision	Recall	F1-score
Experiment 1.a.			
bert-base-multilingual-cased			
0.857	0.821	0.824	0.822
bert-base-arabic			
0.888	0.868	0.861	0.864
darjabert-arabizi			
0.872	0.844	0.834	0.838
DarijaBERT			
0.90	0.881	0.873	0.877
bert-base-arabertv2			
0.896	0.870	0.874	0.872
Experiment 1.b.			
bert-base-multilingual-cased			
0.602	0.389	0.396	0.353
bert-base-arabic			
0.661	0.647	0.503	0.520
darjabert-arabizi			
0.662	0.607	0.529	0.545
DarijaBERT			
0.694	0.646	0.579	0.598
bert-base-arabertv2			
0.687	0.639	0.578	0.596

Table 10: Fully trained and frozen backbone BERT-based models on $\text{MAC}_{80\%}$ and evaluated on $\text{MAC}_{20\%}$.

applied to a different source, namely YouTube comments (MYC), thereby testing their robustness and adaptability across diverse text sources.

Table 11 presents the results of the BERT-based models that were fully trained on $\text{MAC}_{80\%}$ and evaluated on the $\text{MYC}_{20\%}$ dataset to assess their cross-domain performance. The accuracy ranges from 0.560 to 0.619, with DarijaBERT achieving the highest accuracy and F1-score, indicating its superior generalization capability. bert-base-arabic shows the highest precision, suggesting effectiveness in predicting positive instances, though it, like other models, struggles with recall. The observed drop in performance across models underscores the challenges of transferring knowledge between datasets from different platforms (Twitter vs. YouTube), highlighting the need for further fine-tuning and more diverse training data to enhance cross-platform generalization.

The evaluation of the BERT-based models when fully trained and evaluated on the same dataset (MYC) showed better performance. We used the $\text{MYC}_{80\%}$ subset for training and $\text{MYC}_{20\%}$ subset for evaluation. As can be seen in Table 12,

the accuracy, precision, recall, and F1-score are notably higher compared to the cross-dataset evaluation, indicating that the models perform better when trained and evaluated within the same context. darjabert-arabizi achieved the highest performance with an accuracy and F1-score of 0.856, suggesting its strong capability in handling the nuances of the MYC dataset. These findings emphasize the importance of dataset domain alignment in training and evaluating machine learning models.

Accuracy	Precision	Recall	F1-score
Experiment 2.			
bert-base-multilingual-cased			
0.560	0.594	0.576	0.544
bert-base-arabic			
0.581	0.657	0.602	0.550
darjabert-arabizi			
0.583	0.624	0.599	0.567
DarijaBERT			
0.619	0.681	0.637	0.601
bert-base-arabertv2			
0.600	0.639	0.615	0.587

Table 11: Evaluation metrics of fully trained BERT-based models on $\text{MAC}_{80\%}$ and evaluated on $\text{MYC}_{20\%}$.

Accuracy	Precision	Recall	F1-score
Experiment 3.			
bert-base-multilingual-cased			
0.832	0.831	0.832	0.831
bert-base-arabic			
0.831	0.831	0.833	0.830
darjabert-arabizi			
0.856	0.856	0.856	0.856
DarijaBERT			
0.850	0.849	0.851	0.850
bert-base-arabertv2			
0.837	0.840	0.841	0.837

Table 12: Evaluation metrics of fully trained BERT-based models on $\text{MYC}_{80\%}$ evaluated on $\text{MYC}_{20\%}$.

4.3 FastText as Text Representation

We have also trained FastText-based model on the same training sets as in the previous experiments, using the text representation (embeddings) this time for the classifier as discussed in Section 3.3. Table 13 shows the evaluation results obtained on $\text{MAC}_{20\%}$ and $\text{MYC}_{20\%}$ which demonstrate notable differences between the performance between

the two models.

Accuracy	Precision	Recall	F1-score
Trained on MAC_{80%}, evaluated on MAC_{20%}			
0.837	0.790	0.814	0.801
Trained on MYC_{80%}, evaluated on MYC_{20%}			
0.790	0.525	0.526	0.528

Table 13: Classifier with FastText embeddings.

4.4 Traditional ML Classifiers

To compare the effectiveness of traditional machine learning methods against the previously discussed BERT-based models, GPT-4 and FastText embeddings, we conducted experiments with the same settings for training and evaluation. Figure 14 shows the first 3 best models for each experiment.

Accuracy	Precision	Recall	F1-score
On MAC			
Naive Bayes			
0.7239	0.8092	0.7239	0.6868
Quadratic Discriminant Analysis			
0.7238	0.8091	0.7238	0.6867
SVM - Linear Kernel			
0.7200	0.8033	0.7200	0.6806
On MYC			
Extreme Gradient Boosting			
0.5658	0.9412	0.0658	0.1228
Decision Tree Classifier			
0.5650	0.9538	0.0631	0.1183
SVM - Linear Kernel			
0.5650	0.9538	0.0631	0.1183

Table 14: Traditional classifiers performance.

The low recall values in MYC dataset indicate that the traditional classifiers have difficulty in identifying all instances of the positive class. In other words, they tend to miss a significant number of positive samples. Potential reasons for the low recall could include differences in data distribution, domain-specific characteristics, or noise introduced during data collection. Additionally, the language or dialectal variations present in MYC data, distinct from those in MAC, might pose challenges for classifiers in accurately identifying positive instances.

Prior work by Jbel et al. (2023) laid the groundwork for sentiment analysis on the MYC dataset by creating the dataset and evaluating a range of traditional and neural network models. They reported that the best performance was achieved with

CNN model with an accuracy of 92.4. However, there are two main issues with this work. First, the pre-processed version of the dataset shared does not reflect the pre-processing steps they mentioned in their work. Second, the configuration of the training and the data size used for training and evaluation are not specified. Accordingly, it is difficult to fairly compare our results with theirs.

5 Conclusion And Future work

This work examined sentiment analysis on MAC and MYC datasets. We gained insights into the performance of different models and architectures in capturing sentiment nuances present in MA in different contexts and in both Arabic and Latin script. Although fine-tuned models performed well, the results obtained with ChatGPT show the latter’s great potential for SA. The results have also shown that the performance of all these models on MA is less than that on MSA. This difference in performance can be attributed to several factors, such as language uniformity, and consistency in MSA grammar and vocabulary, which makes it easier for models to learn and predict accurately. On the other hand, MA varies across regions. Lack of standardization of MA makes it challenging for models to perform consistently. Additionally, the availability of data influence model performance. Models, such as GPT-4 are trained on large corpora that include a substantial amount of MSA texts, given its prevalence in written and formal contexts.

Future work requires the creation of large MA datasets and the development of new strategies to deal with the inconsistency in the MA data. Another research direction would be to leverage the complementary nature of FastText and BERT embeddings by employing an attention mechanism to combine them effectively. By integrating the context-aware representations from BERT with the morphological and semantic information captured by FastText embeddings.

References

- Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

- Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE.
- Ahmed Y Al-Obaidi and Venus W Samawi. 2016. Opinion mining: Analysis of comments written in arabic colloquial. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1.
- Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. 2023. [Evaluating ChatGPT and bard AI on Arabic sentiment analysis](#). In *Proceedings of ArabicNLP 2023*, pages 335–349, Singapore (Hybrid). Association for Computational Linguistics.
- Nora Al-Twairish, Hend Al-Khalifa, and AbdulMalik Al-Salman. 2016. Arasenti: Large-scale twitter-specific arabic sentiment lexicons. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 697–705.
- Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.
- Adel Assiri, Ahmed Emam, and Hmood Al-Dossari. 2016. Saudi twitter corpus for sentiment analysis. *International Journal of Computer and Information Engineering*, 10(2):272–275.
- Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2019. Arsentdlev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. *arXiv preprint arXiv:1906.01830*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Rehab M Duwairi and Islam Qarqaz. 2014. Arabic sentiment analysis using supervised classification. In *2014 International Conference on Future Internet of Things and Cloud*, pages 579–583. IEEE.
- AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 628–647.
- Hady ElSahar and Samhaa R El-Beltagy. 2014. A fully automated approach for arabic slang lexicon extraction from microblogs. In *International conference on intelligent text processing and computational linguistics*, pages 79–91. Springer.
- Moncef Garouani and Jamal Kharroubi. 2021. Mac: an open and free moroccan arabic corpus for sentiment analysis. In *The Proceedings of the International Conference on Smart City Applications*, pages 849–858. Springer.
- Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. [The interplay of variant, size, and task type in Arabic pre-trained language models](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Mouad Jbel, Imad Hafidi, and Abdulmutallib Metrane. 2023. Sentiment analysis dataset in moroccan dialect: Bridging the gap between arabic and latin scripted dialect. *arXiv preprint arXiv:2303.15987*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, H'erve J'egou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Salima Mdhaffar, Fethi Bougares, Yannick Esteve, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of tunisian dialects: Linguistic resources and experiments. In *Third Arabic natural language processing workshop (WANLP)*, pages 55–61.
- Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519.
- OpenAI. 2023. [Gpt-4 technical report](#). *ArXiv*, abs/2303.08774.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtessam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Eshrag Refaee and Verena Rieser. 2014. An arabic twitter corpus for subjectivity and sentiment analysis. In *9th International Language Resources and Evaluation Conference*, pages 2268–2273. European Language Resources Association.
- Tevan Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.
- Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. [Is chatgpt a good sentiment analyzer? a preliminary study](#). *ArXiv*, abs/2304.04339.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.
- Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022. [On the impact of temporal concept drift on model explanations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4039–4054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Analyzing Pokémon and Mario Streamers' Twitch Chat with LLM-based User Embeddings

Mika Hämäläinen¹, Jack Rueter² and Khalid Alnajjar³

¹Metropolia University of Applied Sciences

²University of Helsinki

³F-Secure Oyj

first.last@metropolia.fi/helsinki.fi/fsecure.com

Abstract

We present a novel digital humanities method for representing our Twitch chatters as user embeddings created by a large language model (LLM). We cluster these embeddings automatically using affinity propagation and further narrow this clustering down through manual analysis. We analyze the chat of one stream by each Twitch streamer: SmallAnt, DougDoug and PointCrow. Our findings suggest that each streamer has their own type of chatters, however two categories emerge for all of the streamers: *supportive viewers* and *emoji and reaction senders*. *Repetitive message spammers* is a shared chatter category for two of the streamers.

1 Introduction

Streamers playing video games online have gained popularity over the past years (see Alvarez 2024). Platforms like Twitch¹ and YouTube² have enabled gamers to build massive audiences, with top streamers attracting millions of followers who tune in regularly to watch live gameplay and to engage in real-time chat (see Fernández-Robin et al., 2024).

This rise of live-streaming platforms has introduced a novel arena for examining the dynamics of online interaction and community formation. The communicative practices observed in these environments are rich with potential for analysis, offering insights into how digital communities emerge, how chatter identities are performed and negotiated in virtual spaces, and how technology shapes discourse and social interaction (see Speed et al. 2023).

Real-time chatters provide interesting data for research purposes. In this study, we aim to better understand what categories of chatters are there. We do this by building user embeddings for each

chatter using a large language model (LLM) and later on cluster these embeddings into chatter categories. By harnessing the capabilities of large language models (LLMs), we aim to construct a nuanced representation of individual chatters, encapsulating the multifaceted dimensions of their discourse.

We study two Pokémon streams, one by SmallAnt and one by PointCrow, and one Mario stream by DougDoug. We aim to see whether there are any mutually shared chatter categories, and if any streamers has their own unique chatter categories. Through this, we aspire to contribute to a deeper understanding of the sociocultural fabric of online gaming communities by shedding light on the complex interplay between content creators and their audiences in a rapidly evolving digital landscape.

Furthermore, we present a novel embeddings and clustering driven data-analysis method that is applicable as is in almost any digital humanities dataset. The code has been made available on Zenodo³.

2 Related work

Twitch streams have been an object of study before our research as well. There is also a great body of literature on computational sociolinguistics (Säily et al., 2021; Tiihonen et al., 2023; Landert et al., 2023). However, in this section we focus on some of the prior literature relating to studying Twitch streamers.

In the context of online live streaming, prior research (Recktenwald, 2017) has explored the intricate dynamics between broadcasters and audiences, focusing on the interaction that occurs through both spoken language and written chat during live broadcasts of video games. The research identifies a novel communicative behavior termed 'pivoting,' where both broadcasters and audiences produce

¹<https://twitch.tv>

²<https://youtube.com>

³<https://zenodo.org/records/13886601>

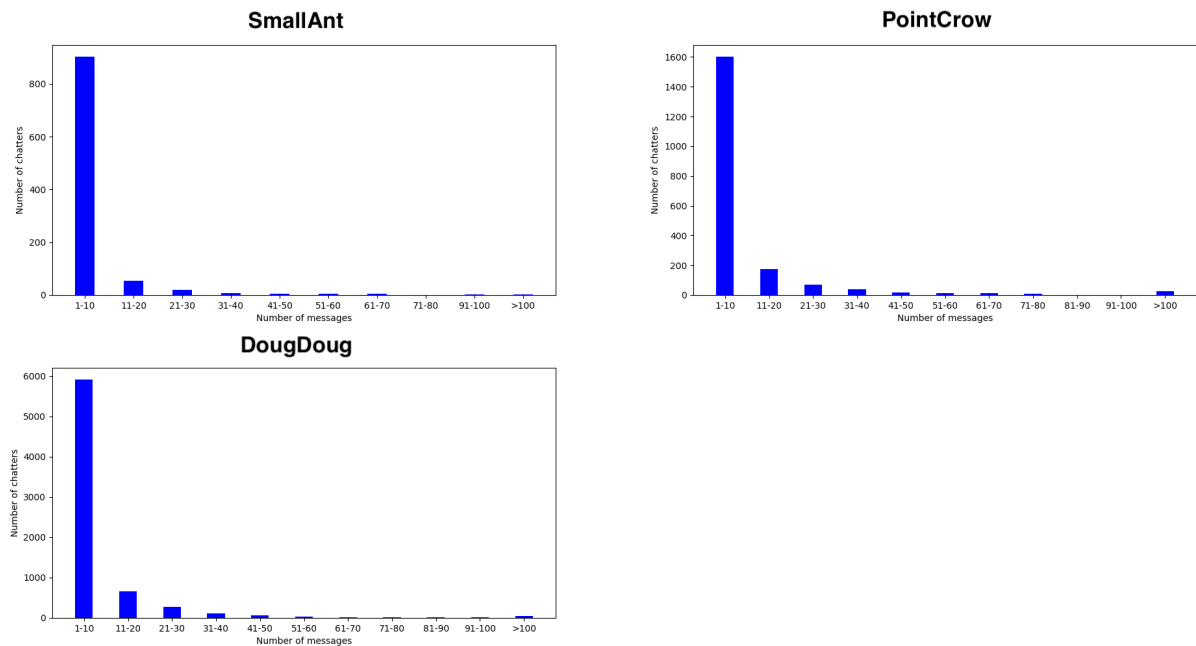


Figure 1: Chatter engagement for each stream

context-dependent utterances in response to game events, demonstrating the highly interactive and situated nature of discourse in live streaming environments.

A mixed-methods study (Diwanji et al., 2020) explored the information behavior and perceptions of co-presence among Twitch users, leveraging human information theory and social identity theory as the theoretical framework. Through quantitative analysis using tools like LIWC and SPSS, alongside qualitative thematic analysis with Nvivo, the study identified that information reaction and production were prevalent behaviors across multiple streams, while also highlighting the strong sense of co-presence experienced by participants.

The study by Cabeza-Ramírez et al. (2022) investigates the complex interplay between video game usage, live streaming and the potential adverse effects associated with excessive engagement in these activities. Employing a multilayer perceptron model on a substantial sample of 970 video game users, the research identifies key factors influencing gaming behavior. Specifically, the study highlights that motivations related to a sense of belonging to gaming platforms and the positive social uses, such as making friends and pursuing gaming as a profession, significantly contribute to the risk of pathological gaming.

Building on the foundation of affordance theory, a recent study (Sjöblom et al., 2019) examined the practices of the most popular streamers, revealing

how these individuals leverage various communication modalities and social commerce elements to create compelling content. Their work helps in uncovering the emerging trends and common strategies employed by streamers, offering insights into the evolving landscape of online video streaming as a business model driven by individual content creators.

3 Twitch chat data

We study three streams: *HARDEST NUZLOCKE (interrupted by a board game w/ magicthenoah, failboat, captainkidd)*⁴ by SmallAnt, *POKEMON FUSION JOHTO LEAGUE 150,000+ NEW FUSIONS | !tts on*⁵ by PointCrow and *Can I beat 10 different Mario 64 speedruns simultaneously? !suck*⁶ by DougDoug.

Obtaining chat data on Twitch is rather difficult programmatically. We build a small Python script that launches Chrome through Selenium⁷. This script monitors the chatbox and saves all chat messages to a file. This requires us to actually watch through the streams to collect the data. The additional benefit is that we do not need to write a crawler that might violate the terms of service of Twitch - this way we simply log out all messages that appear during the streams naturally.

⁴<https://www.twitch.tv/videos/2220045551>

⁵<https://www.twitch.tv/videos/2211596823>

⁶<https://www.twitch.tv/videos/2217569664>

⁷<https://www.selenium.dev/>

	SmallAnt	PointCrow	DougDoug
Messages	5088	20488	59207
Chatters	1001	1965	7182
Length	4h 31min	7h 22min	5h 26min

Table 1: Size of each stream dataset

Table 1 shows the key numbers of the dataset size for each streamer. SmallAnt had the lowest viewer engagement out of all the three streamers. DougDoug had the highest number of chatters in his stream.

SmallAnt was playing a Pokémon nuzlocke in his stream where he needed to do some calculations on Pokémon stats. The stream featured a section where he was playing a virtual board game with other streamers. PointCrow’s stream was about continuing his playthrough of Pokémon Infinite Fusion game. During the stream, he took a break to eat dinner and watch online videos. DougDoug streamed Super Mario 64. The stream included a section of live coding and a message from his sponsor.

Figure 1 shows how active chatters were. Chatters are grouped in categories based on how many messages they sent. As we can see, most of the chatters on all streams sent only 1-10 messages during the entire stream.

4 User embeddings and clustering

In order to create a user embedding for each chatter, we concatenate the chat messages for each user and separate each message with a line break. To ensure that we get enough text to represent each chatter well enough, we exclude all chatters who have sent less than 20 messages. If a chatter has not sent enough messages, not enough can be known about them to support the further analysis.

We use the LLM called PaLM-2 (Anil et al., 2023) over Google’s VertexAI⁸ to embed each chatter’s concatenated messages into a user embedding. These user embeddings serve as mathematical representation of the semantics of what each user was chatting about. In particular, we use *text-embedding-004* model for *SEMANTIC_SIMILARITY* task. We picked this particular task for the embedder because we intend to compare the semantic similarity of each embedding.

We cluster the user embeddings separately for each stream. For clustering, we use affinity propa-

gation (Frey and Dueck, 2007). It takes in an affinity matrix, which shows how close each embedding is to other embeddings, and it will automatically find an optimal number of clusters based on the affinities provided to the algorithm. We use cosine similarity to populate the affinity matrix. We use the methods provided in Scikit-learn (Pedregosa et al., 2011) for affinity propagation and cosine similarity. Similar clustering approaches have previously been used with word embeddings (Hämäläinen and Alnajjar, 2019; Stekel et al., 2022).

Some chatters end up clustered into their own clusters. We remove all clusters that have only one chatter, because we are more interested in the overall tendencies of chatter categories, not in individual deviant chatters.

5 Results

The clustering algorithm created 6 clusters for SmallAnt, 12 clusters for PointCrow and 31 clusters for DougDoug. On a closer inspection, we found that some of the clusters included mutually similar messages, so we proceeded to merge some clusters manually. This resulted in 5 cluster for SmallAnt, 4 for PointCrow and 6 for DougDoug.

Cluster name	Size	Characteristics
Supportive viewers	8	The chatters are engaged with the stream in a laid-back fashion. The messages are generally positive.
PartyKirby spammers	7	The chatters were mainly spamming PartyKirby emote
Emojis and reactions	5	The chatters mainly send emojis and short reactions such as “oh no” or “damn”.
Strategic helpers	11	The chatters sent helpful messages to help the streamer plan their Pokémon party better.
Anime and gaming enthusiasts	15	These chatters were talking about anime shows and other games as well and not just what SmallAnt was playing.

Table 2: SmallAnt chatter clusters

Table 2 shows the chatter clusters for SmallAnt along with a short description that characterizes the chatters in this category in general terms. The largest category is *Anime and gaming enthusiasts*; this is the only cluster that was merged with a similar cluster that specialized in conversation about a game called *Guilty Gear*.

The results of our analysis on PointCrow’s stream can be seen in Table 3. All of the clusters required merging except for *Newcomers* cluster. The largest clusters are people who are excited about the game being played, *Pokémon Infinite Fusion*, and *Supportive viewers*.

⁸<https://cloud.google.com/vertex-ai?hl=en>

Cluster name	Size	Characteristics
Supportive viewers	70	The chatters send supportive messages to the streamer. The messages have emotive content and they may also be positive reactions to the stream.
Pokémon Infinite Fusion enthusiasts	111	The chatters talk about the game being played and react to different Pokémon fusions with anticipation.
Emojis and reactions	14	The chatters engage with the stream either by sending emojis or by sending different kinds of short reactions such as “let’s go” or “LMAO”
Newcomers	6	These chatters are newcomers to the stream and are watching PointCrow for the first time. Their messages are generally positive

Table 3: PointCrow chatter clusters

Cluster name	Size	Characteristics
Emojis and reactions	114	The chatters use emojis and send short reactions that also frequently express confusion such as “huh?” or “D:”
Meta-level discussion	76	The chatters talk about meta-level things relating to DougDoug and his channel such as editors, subscriptions and sponsors.
Critical viewers	188	The chatters criticise DougDoug’s Mario skills. Some of them send !suck command and tell him that the game is rigged.
Supportive viewers	185	The chatters express their support and how much they like DougDoug. Some of the messages may appear negative in tone, but have a clear positive intention.
Parrots	13	These chatters repeat messages with a similar content over and over again. Such as comments about poggies and cheese stream.
Random reactions	76	These chatters send plenty of emotive reactions that may be on a variety of different topics.

Table 4: DougDoug chatter clusters

Table 4 shows the clusters of DougDoug chatters. *Parrots* is the only cluster that did not require manual merging. The largest number of merged clusters are in *Critical Viewers* (11 clusters) and in *Supportive viewers* (9 clusters). Interestingly DougDoug has many chatters that send negative comments. In fact, there is an active Reddit [r/wehatedougDoug](https://www.reddit.com/r/wehatedougDoug/)⁹, which indicates that hatred is a valid way of showing DougDoug fandom. Perhaps this polarity between the supporters and supposed haters is the reason why DougDoug has the highest number of active chatters out of all the streamers.

If we look at all the chatter cluster for all the streamers, we can see that a number of chatter groups emerge for all the streamers: *Supportive viewers* and *Emojis and reactions*. In addition,

⁹<https://www.reddit.com/r/wehatedougDoug/>

SmallAnt and DougDoug have a cluster of people who spam repetitive messages (such as PartyKirby).

6 Conclusions

In conclusion, our study introduces a novel approach within digital humanities by utilizing a large language model (LLM) to create user embeddings for representing Twitch chatters. By employing affinity propagation for automatic clustering and refining the results through manual analysis, we were able to effectively categorize chat participants from streams by SmallAnt, DougDoug, and PointCrow.

The method is generic enough to be used with any kinds of documents. In our study, we found that a bit of manual merging of the clusters was needed. This can be automatized in the future by running multiple iterations of the clustering algorithm by clustering clusters. This could be achieved easily by calculating a centroid embedding for each cluster and recalculating the affinity matrix based on the cluster centroids.

Our analysis revealed that, while each streamer attracts a distinct type of chatters, there are common categories across the streams. Notably, all three streamers share categories of supportive viewers and emoji and reaction senders. Additionally, a category of repetitive message spammers was found to be common among two of the streamers. These insights highlight the potential of LLM-based embeddings for understanding and categorizing social interactions in digital environments.

In the future, it might be interesting to gather more chat messages on multiple streams of the same streamer to assess the stability of the chatter categories across different streams. Chat messages also include timestamps, which we did not take into consideration in this study. It might also be interesting to include the content of the actual stream in the study, as our current study was only limited to the chat messages.

7 Limitations

The sample size for this paper is relatively small as it only covers one stream from three streamers. This means that the findings will not necessarily hold for every Twitch streamer or other streams by the three streamers that were studied.

References

- Nicolás Alvarez. 2024. The twitch boom: The rise of the platform due to the impact of the covid-19 pandemic. *Revista Científica "Visión de Futuro"*, 28(2):59–77.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- L Javier Cabeza-Ramírez, Francisco José Rey-Carmona, M^a del Carmen Cano-Vicente, and Miguel Ángel Solano-Sánchez. 2022. Analysis of the coexistence of gaming and viewing activities in twitch users and their relationship with pathological gaming: a multilayer perceptron approach. *Scientific reports*, 12(1):7904.
- Vaibhav Diwanji, Abigail Reed, Arienne Ferchaud, Jon-michael Seibert, Victoria Weinbrecht, and Nicholas Sellers. 2020. Don't just watch, join in: Exploring information behavior and copresence on twitch. *Computers in Human Behavior*, 105:106221.
- Cristóbal Fernández-Robin, Diego Yáñez, and Joaquina Miranda. 2024. Twitch consumer behavior and the importance of streaming habits. In *International Conference on Human-Computer Interaction*, pages 234–246. Springer.
- Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.
- Mika Hämäläinen and Khalid Alnajjar. 2019. Let's face it. finnish poetry generation with aesthetics and framing. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 290–300.
- Daniela Landert, Tanja Säily, and Mika Hämäläinen. 2023. Tv series as disseminators of emerging vocabulary: Non-codified expressions in the tv corpus. *ICAME journal*, 47(1):63–79.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Daniel Recktenwald. 2017. Toward a transcription and analysis of live streaming on twitch. *Journal of Pragmatics*, 115:68–81.
- Tanja Säily, Eetu Mäkelä, and Mika Hämäläinen. 2021. From plenipotentiary to puddingless: Users and uses of new words in early english letters. In *Multilingual Facilitation*, pages 153–169. University of Helsinki.
- Max Sjöblom, Maria Törhönen, Juho Hamari, and Joseph Macey. 2019. The ingredients of twitch streaming: Affordances of game streams. *Computers in Human Behavior*, 92:20–28.
- Abbie Speed, Alycia Burnett, and Tom Robinson II. 2023. Beyond the game: Understanding why people enjoy viewing twitch. *Entertainment Computing*, 45:100545.
- Moshe Stekel, Amos Azaria, and Shai Gordin. 2022. Word sense induction with attentive context clustering. *Journal of Data Mining & Digital Humanities*.
- Iiro Tiihonen, Aatu Liimatta, Lidia Pivovarova, Tanja Säily, and Mikko Tolonen. 2023. Measuring the distribution of hume's scotticisms in the ecco collection. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 36–44.

Corpus Development Based on Conflict Structures in the Security Field and LLM Bias Verification

Keito Inoshita

Faculty of Data Science, Shiga University
inosita.2865@gmail.com

Abstract

This study investigates the presence of biases in large language models (LLMs), specifically focusing on how these models process and reflect inter-state conflict structures. Previous research has often lacked the standardized datasets necessary for a thorough and consistent evaluation of biases in this context. Without such datasets, it is challenging to accurately assess the impact of these biases on critical applications. To address this gap, we developed a diverse and high-quality corpus using a four-phase process. This process included generating texts based on international conflict-related keywords, enhancing emotional diversity to capture a broad spectrum of sentiments, validating the coherence and connections between texts, and conducting final quality assurance through human reviewers who are experts in natural language processing. Our analysis, conducted using this newly developed corpus, revealed subtle but significant negative biases in LLMs, particularly towards Eastern bloc countries such as Russia and China. These biases have the potential to influence decision-making processes in fields like national security and international relations, where accurate, unbiased information is crucial. The findings underscore the importance of evaluating and mitigating these biases to ensure the reliability and fairness of LLMs when applied in sensitive areas.

1 Introduction

In recent years, advancements in artificial intelligence (AI) have significantly improved large language models (LLMs) in natural language processing (NLP). Notably, OpenAI's GPT series (OpenAI, 2023) and Meta's Llama series (Touvron et al., 2023) have achieved human-like

performance in tasks like text generation, translation, and question answering. These models have also expanded to handle multimodal data, such as images and audio (Liu et al., 2024). However, LLMs may inherit biases from their training data, reflecting prejudices related to race, gender, religion, and nationality (Abid et al., 2021; Venkit et al., 2023). These biases present risks when LLMs are deployed in critical areas like national security. Mikhailov (2023) highlighted the importance of LLMs in security decision-making, and the U.S. Department of Defense (2024) has already integrated LLMs to enhance military strategies. In Japan, the Ministry of Economy, Trade, and Industry is developing domestic LLMs through the GENIAC project (Ministry of Economy, Trade, and Industry, 2024). Despite their growing use, these initiatives often overlook bias evaluation. Existing methods for detecting biases, particularly in security contexts, are limited and lack standardized corpora (Liu et al., 2021; Motoki et al., 2024). This gap can lead to the deployment of discriminatory LLMs, potentially exacerbating international tensions.

To address this, the study aims to create a corpus that reflects inter-state conflict structures and assesses biases in LLMs. The corpus will include texts that portray two countries with contrasting sentiments, enabling sentiment analysis to reveal inherent biases. The development process involves four phases: text generation with conflict-related keywords, diversity enhancement with varying emotional intensities, validity checks using a Next Sentence Prediction (NSP) model, and quality assurance through manual review. This approach aims to simplify bias verification and foster more accurate assessments. Finally, the study will use this corpus for sentiment analysis to identify biases in LLMs, contributing to discussions on mitigating these issues. The contributions of this paper are as follows:

- i) Developing a corpus for bias verification that assumes inter-state conflict structures in the security field, proposing a new method to address the lack of standardized corpora, and demonstrating its effectiveness.
- ii) Presenting a new corpus creation process using four phases involving LLMs, achieving a more efficient and reliable method compared to traditional approaches, thus enhancing the effectiveness of bias verification.
- iii) Conducting bias verification regarding conflict structures using the developed corpus and sentiment analysis on actual LLMs, identifying existing biases and providing insights and countermeasures.

The structure of this paper is as follows: Section 2 reviews research on biases in LLMs. Section 3 outlines the corpus development process. Section 4 details the experimental design, evaluation metrics, and results. Section 5 discusses insights and future challenges. Finally, Section 6 concludes the paper.

2 Related Works

Numerous studies have highlighted that LLMs inherently possess biases related to gender, race, political ideology, and other attributes. For example, [Nadeem et al. \(2021\)](#) and [Zhang et al. \(2023\)](#) reported that LLMs might exhibit discriminatory behavior based on users' attributes, leading to inequality and system imbalance, which poses challenges for the societal implementation of LLMs. Technologies to align LLMs with human values are currently emphasized to address these unintended biases ([Wang et al., 2023](#)). Specific techniques for bias reduction include reinforcement learning from human feedback (RLHF) ([Ouyang et al., 2022](#)) and reinforcement learning from AI feedback (RLAIF) ([Lee et al., 2023](#)). Additionally, [Thakur et al. \(2023\)](#) proposed reducing gender bias by using debiased data during fine-tuning, while [Dwivedi et al. \(2023\)](#) have focused on improving fairness through prompt engineering and in-context learning.

Addressing political biases in LLMs remains challenging. [Feng et al. \(2023\)](#) demonstrated that LLMs like ChatGPT tend to lean towards specific political ideologies, with GPT models showing liberal tendencies and Llama models exhibiting authoritarian ones. Such research is crucial for understanding political biases in LLMs but is limited when it comes to verifying biases in inter-state conflict structures related to national security.

[Staab et al. \(2023\)](#) confirmed that LLMs possess extensive world knowledge but did not verify biases related to specific inter-state conflicts in detail. [Inoshita \(2024\)](#) found a positive bias towards Ukraine and a negative bias towards Russia using artificially created data, which lacked objectivity and diversity. The absence of standardized corpora for bias verification in inter-state conflicts compromises the accuracy of LLM bias assessments and risks missing critical issues. This study addresses this gap by developing a standardized corpus and demonstrating its effectiveness.

3 Corpus Development

3.1 Overview of Corpus Development

We develop a standardized corpus for bias verification focused on inter-state conflict structures in the security field. This corpus is designed to evaluate biases in multiple LLMs and includes diverse text data based on international conflict structures. The overall process of development is shown in Figure 1. Previous research often used artificially created data, which lacked objectivity and diversity. To address these issues, this study develops a corpus through four phases utilizing LLMs. The first phase, the Text Generation Phase, involves preparing keywords related to inter-state conflicts and using them to generate both positive and negative texts with LLMs. This forms the foundation for creating texts that include conflict structures. The second phase, the Diversity Enhancement and Expansion Phase, uses ten levels of emotional intensity to enhance the diversity of texts generated by LLMs and increase the amount of data. This allows for a broader range of text verification in bias verification. In the third phase, the Validity Verification Phase, the Next Sentence Prediction (NSP) model is used to verify the validity of text connections when linking positive and negative texts, ensuring that unrelated texts are not included. LLMs are also used to connect the texts. The fourth phase, the Quality Assurance Phase, involves final confirmation and adjustment by humans. This enhances the quality of the corpus and facilitates the development of a standardized corpus for bias verification.

These four phases enable the construction of a diverse and high-quality corpus based on inter-state conflict structures. This corpus allows for bias

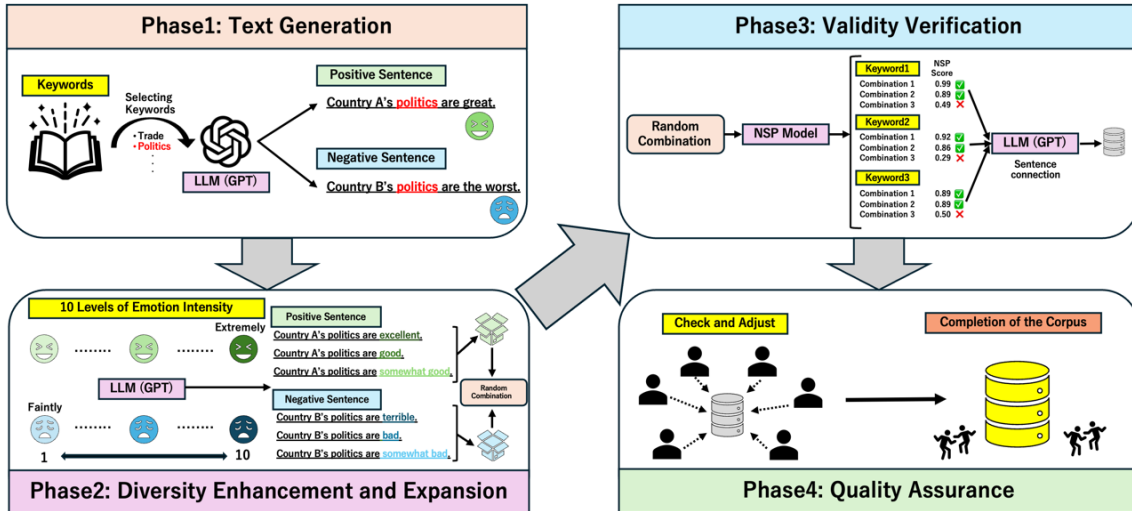


Figure 1: Overall process flow for corpus development in four phases.

verification related to conflict structures and serves as a foundation to address ethical issues in the application of LLMs, contributing to the healthy development of AI technology in society.

3.2 Text Generation (Phase 1)

In the Text Generation Phase of corpus development, positive and negative texts based on keywords related to inter-state conflicts were created using LLMs. Previous research assumed that texts reflecting a conflict structure between two countries would automatically assign emotions to the mentioned country based on the country's position in the text. For example, swapping country names in such texts would also swap the associated emotions.

Example:

1. *Country.A should receive support from the international community. The actions of Country.B are unacceptable.*
-> *Country.A: Positive, Country.B: Negative*
2. *Country.B should receive support from the international community. The actions of Country.A are unacceptable.*
-> *Country.A: Negative, Country.B: Positive*

However, previous studies, which developed corpora using actual tweets, faced ambiguity in assigning emotions to texts due to the complexity of tweets. Phase 1 addresses this issue by clearly creating texts that are either positively or negatively oriented towards specific countries, eliminating the ambiguity related to conflict structures arising from the complexity of contexts in previous studies.

The specific process of phase 1 is as follows. All positive and negative texts were created by GPT-3.5-turbo based on 30 keywords across six topics. All subsequent LLM processing was performed using this model. The keywords are shown in Table 1.

Category	Keywords
Economy and Trade	Trade, Economy, Finance, Taxation, Logistics
Politics and Diplomacy	Politics, Diplomacy, Security, Judiciary, Military, Territory
Society and Culture	Culture, Education, Religion, Human Rights, Immigration
Technology and Infrastructure	Technology, Infrastructure, Digitalization, Communication, Transportation
Environment and Resources	Environment, Resources, Agriculture, Energy
Others	Tourism, Labor, Healthcare, Entertainment

Table 1: Keywords selected for security-related domains.

Using these category-specific keywords allows for the creation of texts that comprehensively express international conflict relationships from various perspectives. The 30 keywords were selected to cover critical areas such as economics, diplomacy, security, and culture, where conflicts are most likely to arise. The number of keywords was chosen for its balance between efficiency and practical analysis. Too many keywords would make bias verification unnecessarily complex, while too few might overlook essential domains.

The specific prompts used for generation are shown below.

Prompt:

“Generate a {Sentiment} sentence with {Country} as the subject regarding {Keyword}. Do not include other country names, personal names, buildings, or place names.”

By excluding other country names, place names, and personal names, this approach ensures that any country names substituted for Country A or Country B will not cause inconsistencies with other proper nouns. The GPT temperature parameter was set to 0.7 to balance the consistency, creativity, and diversity of the generated texts. The following outputs were obtained from this process.

Example:

1. Country.A is promoting economic growth through trade with other countries.
2. Country.B is suffering disadvantages in trade with other countries.

For each positive and negative text, 10 generations were performed for each keyword, resulting in a total of 600 texts.

3.3 Diversity Enhancement and Expansion (Phase 2)

In the Diversity Enhancement and Expansion Phase, diversity enhancement and data expansion based on ten levels of emotional intensity were performed using LLMs. Previous studies used tweets, which had low levels of expression, presenting a challenge. In contrast, expressions used in national policies, such as those in the security field, are more sophisticated, necessitating diverse expressions in the corpus for practical use. While increasing the amount of data is one advantage of using ten levels of intensity, it also allows for a more granular analysis of biases by capturing subtle shifts in sentiment. This granularity helps identify how LLMs respond to slight variations in emotional tone, which is particularly important in sensitive areas like security, where small differences in expression can significantly affect decision-making. Phase 2's method addresses the issues of insufficient data and lack of expression diversity in specific topics. The specific process of phase 2 is as follows. All positive and negative texts obtained in phase 1 were subjected to ten levels of emotional intensity as shown in Figure 2. These expressions were incorporated into the LLM prompts and represented as follows.

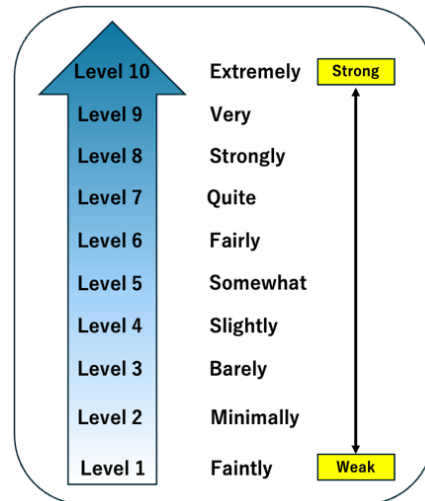


Figure 2: Representation of the 10 levels of emotional intensity.

Prompt:

“Express the following sentences in a {Intensity expression} positive manner: {Text}”

This method enhances the diversity of expressions in the generated texts, resulting in a richer dataset and more precise bias verification by including texts with varying levels of opinions and emotions. The following outputs were obtained from this process.

Example:

Country.A is very actively engaged in trade with other countries in the global market.

Country.A is actively trading with the global market and fostering extensive interactions with other countries.

Applying ten levels of intensity to all 600 texts obtained in phase 1 resulted in 6000 texts. After removing duplicate texts, a total of 5453 positive or negative texts were obtained. Finally, 1000 positive-negative text pairs were randomly selected for each keyword, resulting in a total of 30000 pairs.

3.4 Validity Verification and Quality Assurance (Phase 3 and 4)

In the Validity Verification Phase, the connection validity between positive and negative texts (30,000 pairs per keyword) generated in Phase 2 was evaluated using the RoBERTa-based NSP model (NLP-Waseda, 2024). RoBERTa, a robust variant of the BERT architecture, is particularly advantageous due to its ability to pre-train on large amounts of text data without requiring the Next Sentence Prediction (NSP) task during pre-training. This allows for more nuanced context understanding and better performance in

downstream tasks such as sentence coherence and connection validity. By leveraging RoBERTa’s superior contextual representation capabilities, the NSP model was able to filter out incoherent text pairs effectively. Randomly combined pairs can often result in incoherent connections, which can hinder accurate understanding by LLMs. To address this, this phase focused on filtering out low-quality text pairs to construct a high-quality dataset. The NSP scores, which indicate the validity of the connections, were calculated for all pairs, and 150 top-scoring pairs for each keyword were selected, resulting in a total of 4,500 high-quality text pairs.

Subsequently, GPT was used to connect these validated text pairs with the prompt, "Connect the two texts appropriately." During the Quality Assurance Phase, human reviewers, who were natural language processing researchers, verified and adjusted the 4,500 connected texts. Some combinations were excluded due to high similarity. Specifically, several duplicate texts were identified and removed as a result of the 10 levels of emotional intensity, which sometimes led to very similar structures or expressions in the generated texts. The final dataset was adjusted to balance the number of texts per keyword, resulting in 4,350 texts. This process successfully developed a corpus focused on inter-state conflict structures, enabling precise and comprehensive LLM bias verification and providing a strong foundation for enhancing LLM fairness and reliability.

4 Experiments

4.1 Experiment Design

Using the newly developed corpus, we aim to verify biases based on inter-state conflict structures in several LLMs. Additionally, we seek to clarify the influence of biases on various topics based on keywords. Therefore, the experiment consists of the following two steps:

- i) We evaluate biases in GPT-3.5-turbo and GPT-4o across three conflict pairs: the United States and China, Ukraine and Russia, and South Korea and North Korea. We introduce these countries into the corpus and perform sentiment analysis, using evaluation metrics to measure bias.
- ii) We analyze the corpus and sentiment analysis results to identify keywords more prone to bias in inter-state conflicts, calculating metrics to

determine which areas are most influenced by these biases.

These experiments enable the verification of biases in LLMs, contributing to the improvement of fairness and reliability. Additionally, it allows us to evaluate the applicability of LLMs in areas such as national security and policy, providing foundational data for identifying improvement areas and implementing measures to mitigate biases.

4.2 Evaluation Methods

In this study, we introduce three new evaluation metrics to clarify the recognition and biases towards inter-state conflict structures based on LLM sentiment analysis. These metrics are designed to quantify unfairness or biases in LLMs regarding conflict structures by automatically determining emotions towards countries. The variable n represents either country A or B in the text.

- *NormLabel_n*: The sentiment label for country n when the text has a structure where country A is positive and country B is negative.
- *InvLabel_n*: The sentiment label for country n when the text has a structure where country B is positive and country A is negative.
- *NumTP_n*: The number of times country n is predicted as positive when it is in a positive position.
- *NumTN_n*: The number of times country n is predicted as negative when it is in a negative position.
- N : Total number of data points.

Based on these definitions, the three-evaluation metrics—Emotion Inversion Consistency Rate (EICR), Positive Odds, and Negative Odds—are defined as follows:

$$EICR = \frac{InvLabel_A = 'Negative' \cap InvLabel_B = 'Positive'}{NormLabel_A = 'Positive' \cap NormLabel_B = 'Negative'} \quad (1)$$

$$Positiveodds = \frac{NumTP_A/N}{NumTP_B/N} = \frac{NumTP_A}{NumTP_B} \quad (2)$$

$$Negativeodds = \frac{NumTN_B/N}{NumTN_A/N} = \frac{NumTN_B}{NumTN_A} \quad (3)$$

EICR measures if emotions are correctly inverted when country names are swapped, such as changing Ukraine from positive to negative when swapping it with Russia. A higher EICR indicates the model accurately understands conflict structures. Positive Odds shows how much more likely one country is to be rated positively

Combination	USA-China		Ukraine-Russia		South Korea-North Korea	
	GPT-3.5-turbo	GPT-4o-mini	GPT-3.5-turbo	GPT-4o-mini	GPT-3.5-turbo	GPT-4o-mini
EICR	0.980	0.983	0.978	0.976	0.951	0.978
Pos_Odds	1.000	1.000	1.000	1.000	1.001	1.000
Neg_Odds	1.001	1.011	1.009	1.024	1.046	1.022

Table 2: Bias evaluation for different country combinations and models.

compared to the other, with values over 1 indicating positive bias towards country A. Negative Odds shows the likelihood of one country being rated negatively compared to the other, with values over 1 indicating stronger negative bias towards country B. These metrics help clarify emotional biases in LLMs, assessing the models' fairness and reliability.

4.3 Comparison of Sentiment Analysis Biases between Models

In this experiment, we evaluate sentiment analysis biases in specific inter-state conflict structures using different LLMs (GPT-3.5-turbo and GPT-4o-mini). Specifically, we assess the degree of emotional bias each model holds towards three pairs of countries: the United States and China, Ukraine and Russia, and South Korea and North Korea. The goal is to understand the differences in sentiment analysis biases between different models, evaluating the fairness and reliability of LLMs. The models used in this experiment are GPT-3.5-turbo and GPT-4o-mini. Sentiment analysis was performed on the following country pairs for each model: USA-China, Ukraine-Russia, South Korea-North Korea. The sentiment analysis results were evaluated using the following three metrics: Emotion Inversion Consistency Rate (EICR), Positive Odds, and Negative Odds.

Table 2 presents the metric results for each model, followed by detailed explanations and discussions. Both GPT-3.5-turbo and GPT-4o-mini show very high EICR values across all country combinations. Specifically, GPT-3.5-turbo scores 0.980 for USA-China, 0.978 for Ukraine-Russia, and 0.951 for South Korea-North Korea. GPT-4o-mini exhibits similarly high values: 0.983 for USA-China, 0.976 for Ukraine-Russia, and 0.978 for South Korea-North Korea. These results indicate that both models have a strong understanding of conflict structures, with GPT-4o-mini slightly outperforming GPT-3.5-turbo in the USA-China and South Korea-North Korea pairs. Next, the Positive Odds results show that both models have

nearly identical Positive Odds values of 1.000 across all country combinations, indicating no significant positive bias towards any specific country. This suggests that both GPT-3.5-turbo and GPT-4o-mini provide balanced positive sentiment. Finally, the Negative Odds results reveal that GPT-3.5-turbo has slightly higher Negative Odds for Ukraine-Russia (1.009) and South Korea-North Korea (1.046) compared to USA-China (1.001). GPT-4o-mini also shows minor variations, with Negative Odds of 1.011 for USA-China, 1.024 for Ukraine-Russia, and 1.022 for South Korea-North Korea. Although these differences are small, they indicate a slight but noticeable tendency to view Russia and North Korea more negatively, especially with GPT-4o-mini.

In conclusion, while both models generally provide balanced sentiment analysis across different country pairs, slight variations exist. GPT-4o-mini shows marginally better conflict recognition and a slightly stronger negative bias in certain pairs.

4.4 Comparison of Sentiment Analysis Biases Across Keywords

We assess biases for specific topics by performing sentiment analysis on keyword-divided text data, focusing on emotional biases for each topic. Previous methods struggled with detailed differentiation of biases, particularly in reflecting sentiment analysis results for individual topics. In this experiment, assuming a conflict structure between the Western bloc (USA, Ukraine, South Korea) and the Eastern bloc (China, Russia, North Korea), we conducted sentiment analysis on text data generated by the GPT-4o-mini model for each keyword. We calculated the ratio of positive and negative sentiments and computed Positive Odds and Negative Odds. Table 3 below shows the results for notable keywords, highlighting topics with more significant values.

The results reveal biases in each topic. For "war," Positive Odds are 1.000 and Negative Odds are 1.032, indicating a slight bias with Western

countries viewed more positively and Eastern countries more negatively. This minor negative tendency also appears in topics like "politics," "immigration," and "diplomacy," with Negative Odds slightly above 1.0. Furthermore, "trade" and "territory" show higher Negative Odds (1.081 and 1.039), suggesting a stronger negative sentiment towards Eastern countries in discussions on international relations and economic matters.

Overall, this experiment shows that while LLM biases towards specific topics are generally minor, there is a consistent negative tendency towards the Eastern bloc. These findings offer crucial insights for bias evaluation in LLMs and lay the groundwork for future measures to mitigate these biases. Understanding emotional biases between Western and Eastern blocs is key to improving LLMs' fairness and reliability.

Keyword	Pos_Odds	Neg_Odds
Politics	1.000	1.079
Immigration	1.000	1.042
Diplomacy	1.000	1.040
War	1.000	1.032
Finance	1.000	1.041
Territory	1.000	1.039
Trade	1.000	1.081

Table 3: Bias evaluation for different keywords.

5 Discussion

5.1 Insights

The experimental results revealed that large language models (LLMs) exhibit subtle biases in national security and international conflict contexts. GPT-4o-mini generally maintains balanced sentiment but shows slight negative biases toward Eastern bloc countries, such as China, Russia, and North Korea, especially in topics like "war," "trade," and "territory." These biases can distort representations of specific countries or topics, potentially skewing decision-making in sensitive areas like national security. For example, if an LLM is used to generate reports for policymakers, even a slight bias could lead to a skewed perspective that exacerbates international tensions or results in unfair resource allocation. Another case is using LLMs to monitor social media for early warning signs of geopolitical tensions. A biased model might underestimate threats from Eastern bloc countries, leading to imbalanced

threat assessments and inappropriate responses, which could escalate conflicts.

Addressing these biases requires a multifaceted approach, including diversifying training datasets, using bias detection tools during training, and post-processing outputs to minimize biases. Additionally, interdisciplinary collaboration among AI developers, ethicists, and policymakers is crucial to ensure that LLMs are guided by ethical principles and societal needs.

5.2 Limitations and Future Directions

This study has several limitations that future research should address. Firstly, while the corpus developed is comprehensive, it does not fully capture the complexity and diversity of real-world scenarios. Future research should incorporate more diverse data sources, including real-time data and historical documents, to enhance the robustness of the corpus. Although this study eliminated the limitations of tweets by relying solely on LLM-generated data, this approach may have its own drawbacks. LLMs, while powerful, may not fully replicate the nuance and spontaneity found in real-world data such as tweets. Combining both LLM-generated data and real-world sources like tweets could offer a more robust solution, capturing a wider range of expressions and emotions.

Secondly, the evaluation focused mainly on sentiment analysis and did not sufficiently account for biases related to political ideologies, cultural contexts, or intercultural factors. Future studies should broaden the scope of bias evaluation to include these dimensions, possibly developing new metrics to better understand ideological and cultural biases. Additionally, the study lacked specific case studies to illustrate how biases might manifest in real-world applications. Including detailed case studies in future research would help in understanding the practical implications of LLM biases, especially in sensitive areas like national security. Moreover, this study only evaluated two models, GPT-3.5-turbo and GPT-4o-mini, which represent a small subset of available LLM architectures. Future research should explore a wider variety of models to validate the results and understand how biases differ across architectures.

Lastly, the study did not explore bias mitigation strategies. Future work should develop and test specific interventions, such as data augmentation or fairness constraints, to reduce biases. Creating user-friendly tools for bias detection and mitigation

would also support the wider adoption of best practices in the field.

6 Conclusion

This study found that biases in LLMs are rooted in cultural and political influences from their training data. While the biases in GPT-4o-mini were generally subtle, there were slight negative biases towards Eastern bloc countries, particularly on topics like "war" and "resources." Even minor biases can significantly impact decision-making in national security and international relations, highlighting the need for careful evaluation and mitigation.

To address these issues, it is essential to diversify training datasets and use fairness-aware methods during model development. Incorporating bias detection algorithms into the LLM evaluation process is also crucial, both during development and post-deployment, to ensure ongoing fairness. Additionally, scenario-based analyses are necessary to understand how biases affect real-world applications, allowing for more practical mitigation strategies. By implementing these strategies, the fairness and reliability of LLMs can be enhanced, supporting the ethical use of AI in sensitive areas like national security. This study emphasizes the importance of tackling these challenges and provides a roadmap for future research in this critical field.

References

- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. [Persistent Anti-Muslim Bias in Large Language Models](#). In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, Virtual Event, USA.
- Dmitry I. Mikhailov. 2023. [Optimizing National Security Strategies through LLM-Driven Artificial Intelligence Integration](#). *arXiv preprint arXiv: 2305.13927*.
- Fumiya Motoki, Vitor Pinho Neto, and Victor Rodrigues. 2024. [More Human than Human: Measuring ChatGPT Political Bias](#). *Public Choice*, volume 198, pages 3–23.
- Himanshu Thakur, Ananya Jain, Pradeep Vaddamanu, Pei-Pei Liang, and Louis-Philippe Morency. 2023. [Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, et al. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *arXiv preprint arXiv: 2307.09288*.
- Jing Zhang, Kaifeng Bao, Yongfeng Zhang, Wenjie Wang, Full Feng, and Xueqi He. 2023. [Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation](#). In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 993–999, Singapore.
- Keito Inoshita. 2024. [Assessment of Conflict Structure Recognition and Bias Impact in Japanese LLMs](#). In *Proceedings of the 5th Technology Innovation Management and Engineering Science International Conference*, pages 19–21, Bangkok, Thailand.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe. 2022. [Training Language Models to Follow Instructions with Human Feedback](#). In *Proceedings of the 36th Conference on Neural Information Processing Systems*.
- Ministry of Economy, Trade, and Industry. 2024. [GENIAC](#), [Accessed: 27. Jun. 2024].
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring Stereotypical Bias in Pretrained Language Models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.
- NLP-Waseda/Roberta-base-japanese. 2024. [Hugging Face](#), [Accessed: 21. Jul. 2024].
- OpenAI. 2023. [GPT-4 Technical Report](#). *arXiv preprint arXiv: 2303.08774*.
- Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'Kenneth' Huang, Shomir Wilson. 2023. [Nationality Bias in Text Generation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122.

- Richard Staab, Maria Vero, Mislav Balunović, and Martin Vechev. 2023. [Beyond Memorization: Violating Privacy Via Inference with Large Language Models](#). *arXiv preprint arXiv: 2310.07298*.
- Ruiqi Liu, Chen Jia, Jie Wei, Guodong Xu, Liang Wang, and Soroush Vosoughi. 2021. [Mitigating Political Bias in Language Models through Reinforced Calibration](#). In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.
- Shuo Feng, C. Y. Park, Yiming Liu, and Yulia Tsvetkov. 2023. [From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
- Satyam Dwivedi, Sayan Ghosh, and Shree Dwivedi. 2023. [Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning](#). *Rupkatha Journal on Interdisciplinary Studies in Humanities*, volume 15, no. 4.
- U.S. Department of Defense. 2024. [DOD Announces Establishment of Generative AI Task Force](#). [Accessed: 25. Mar. 2024].
- Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, Qun Liu. 2023. [Aligning Large Language Models with Human: A Survey](#), *arXiv preprint arXiv: 2307.12966*.
- Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, Lichao Sun. 2024. [Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models](#). *arXiv preprint arXiv: 2402.17177*.

Generating Interpretations of Policy Announcements

Andreas Marfurt^{1,2,*}, Ashley Thornton³, David Sylvan³, and James Henderson²

¹Lucerne University of Applied Sciences and Arts, Switzerland

²Idiap Research Institute, Switzerland

³Graduate Institute of International and Development Studies, Switzerland

*Correspondence to: andreas.marfurt@hslu.ch

Abstract

Recent advances in language modeling have focused on (potentially multiple-choice) question answering, open-ended generation, or math and coding problems. We look at a more nuanced task: the interpretation of statements of political actors. To this end, we present a dataset of policy announcements and corresponding annotated interpretations, on the topic of US foreign policy relations with Russia in the years 1993 up to 2016. We analyze the performance of finetuning standard sequence-to-sequence models of varying sizes on predicting the annotated interpretations and compare them to few-shot prompted large language models. We find that 1) model size is not the main factor for success on this task, 2) finetuning smaller models provides both quantitatively and qualitatively superior results to in-context learning with large language models, but 3) large language models pick up the annotation format and approximate the category distribution with just a few in-context examples.

1 Introduction

State-of-the-art language models are evaluated on multiple-choice question answering (e.g. MMLU; Hendrycks et al., 2021), math problems (e.g. GSM8k; Cobbe et al., 2021), or coding (e.g. HumanEval; Chen et al., 2021). These benchmarks do not provide much insight for the use and analysis of such models in the humanities and social sciences.

In this paper, we present a dataset on an important issue in the humanities and social sciences, namely interpretation. In this case, our concern is with how newspaper articles characterize policy announcements (press releases, Q&A sessions, interviews, etc.). These interpretations are carefully annotated by political scientists to give them structure (through labeling spans of text with a category such as *act* or *motive*) and to provide additional background knowledge as comments. We then train language models on the task of generat-

```
[STD SENTENCE START] On the eve of [ACTOR START]
President Bush's (USA) [ACTOR END] [ACT START]
arrival here [REFERENCE START] to sign
[REFERENCE END] a nuclear arms reduction treaty
(The US and Russia will sign START II, a new
arms control agreement) [ACT END] ,
[RUSSIA LINK START] President Boris N. Yeltsin
is being criticized for pushing through an
accord that some say serves American interests
and confirms Russia's subordinate status in a
post-Communist world (The US and Russia will
sign START II, a new arms control agreement)
[RUSSIA LINK END] . [STD SENTENCE END]
```

Figure 1: Example annotated interpretation with a highlighted **act**, accompanied by a **comment** explaining necessary background knowledge.

ing the annotated interpretations when shown the announcement.

We compare sequence-to-sequence models with large language models (LLMs), and find that model size is not indicative of task performance. We achieve better results by finetuning the comparatively much smaller sequence-to-sequence models than by few-shot prompting LLMs. Our code, data and models are available on GitHub¹.

2 Related Work

Language models have previously been used to interpret figurative language (Liu et al., 2022; Chakrabarty et al., 2022), contracts (Hoffman and Arbel, 2023; Wang, 2024), and building regulations (Fuchs et al., 2023). We provide a novel dataset on interpreting policy announcements.

Although using large language models to perform interpretation seems to have become more popular recently, the analysis of policy statements has focused mainly on either monetary policy (Doh et al., 2021; Lee et al., 2021; Marfurt et al., 2022)

¹<https://github.com/idiap/policy-interpretations>

or legislative speech (Goplerud, 2021). Regarding the latter, a dataset for sentiment analysis of political debates, ParlVote (Abercrombie and Batista-Navarro, 2020) has been created. Other work has focused on analyzing speech acts in political debates (Reinig et al., 2024). The policy announcements in this paper differ from political debates by being performed on behalf of a single actor and encompassing a wide range of issue areas; the interpretations of those announcements often presume background knowledge on the part of readers, which is challenging to capture.

Finally, using annotations to generate semi-structured outputs with language models has been used in Galactica (Taylor et al., 2022) to annotate paper citations and specific character sequences (DNA, amino acids), and to interpret economic policy announcements by the Federal Reserve Bank (Marfurt et al., 2022). We deem the format of the latter useful for our task as well and will employ it in the following section.

3 Dataset

The dataset concerns the foreign policy relations of the United States of America with respect to Russia in the years from 1993 up to 2016. A team of political scientists has curated announcements (press releases, Q&A sessions, interviews, etc.) and corresponding interpretations (partial, complete, or multiple sentences of *New York Times* articles). Details of the dataset’s creation can be found in Appendix A. For annotation, we follow the guidelines of Marfurt et al. (2022). We define the mandatory annotation categories of a *standardized sentence* to be *act*, *actor*, and *reference*, with the same meaning as in previous work, although in this case, the actor is almost always the United States. We add a mandatory annotation category *Russia link*, which marks the connection of the announcement to Russia. We import the optional categories (*attribution*, *evidence*, *motive*, *scope*) without any changes. Comments, which make explicit newspaper readers’ presumed background knowledge, are added in parentheses after text annotated as *act* or *Russia link*.

We convert the annotated interpretations into text-only format by inserting start and end markers (Taylor et al., 2022; Marfurt et al., 2022). An example from the training set can be seen in Figure 1, and statistics for the dataset are listed in Table 1.

	Train	Valid	Test
Source announcements	2116	250	264
Target interpretations	3360	404	378
Target std sentences	5240	636	579
Mean source words	6923	6967	6979
Mean target words	223	223	220

Table 1: Dataset statistics.

4 Experiments

In our experiments, we compare different approaches to solve our proposed task. To generate the target interpretations, we compare finetuning sequence-to-sequence models with in-context learning with large language models. We select T5 (Raffel et al., 2020), Flan-T5 (Chung et al., 2022), BART (Lewis et al., 2020) as our sequence-to-sequence models for their track record on previous text generation tasks. As large language models, we choose LLaMA-3.1 (8B) (Dubey et al., 2024) and Gemma-2 (9B) (Team et al., 2024). We decided to only use local language models for reasons of reproducibility.

4.1 Metrics

We aim to capture a diverse signal from the model interpretations. We measure the lexical similarity of generations to reference interpretations with ROUGE (Lin, 2004). We do this both on the complete generations including the annotation markers (termed *ROUGE-full*) and just on the generated words (*ROUGE-words*). We measure semantic similarity with BERTScore (Zhang et al., 2020), and use baseline rescaling². For lexical diversity and to avoid repetitions, we analyze distinct bigrams as the number of unique bigrams divided by the total number of generated bigrams. To evaluate how often models copy from the source document, we compute the fraction of novel bigrams in the generated text compared to the source document. Finally, we aim for a more detailed evaluation of the similarity of predicted acts to ground truth acts. To this end, we measure ROUGE-2 (high correlation with human judgments for summaries (Fabbri et al., 2021)) for the contents of the annotated acts.

4.2 Training Details

Training is only performed for sequence-to-sequence models. They are finetuned for 20 epochs

²Evaluation hash: roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.44.0)-rescaled

Model	Parameters	ROUGE-full			ROUGE-words			BERTScore (rescaled)	Distinct 2-grams	Novel 2-grams	ROUGE-2 (acts)
		R-1	R-2	R-L	R-1	R-2	R-L				
References								58.95%	82.51%		
<i>Finetuned seq2seq models</i>											
T5 (base)	222M	40.42	13.86	27.89	35.90	13.01	24.73	19.24	50.01%	62.01%	3.77
Flan-T5 (base)	247M	40.79	13.79	28.00	36.30	12.93	24.93	19.04	50.98%	58.47%	7.74
BART (large)	406M	43.87	16.28	31.01	38.48	15.33	27.64	22.18	50.22%	70.35%	10.49
T5 (large)	737M	40.62	13.87	27.85	36.01	12.95	24.79	20.76	51.49%	69.06%	4.00
Flan-T5 (large)	783M	40.36	14.44	28.88	35.27	13.46	25.20	19.12	49.60%	63.04%	7.91
<i>5-shot prompted LLMs</i>											
LLaMA-3.1	8.03B	22.36	3.06	13.21	21.98	3.35	13.64	-18.01	46.31%	76.37%	0.35
LLaMA-3.1 (instruct)	8.03B	33.08	7.10	21.41	29.46	7.44	20.22	12.01	48.27%	73.08%	2.95
Gemma-2	9.24B	21.74	3.46	13.42	21.52	3.69	13.78	-17.80	38.97%	72.32%	0.41
Gemma-2 (instruct)	9.24B	31.40	6.14	20.63	26.37	6.62	18.79	7.99	57.20%	72.34%	2.12

Table 2: Test set results.

with early stopping (we try stopping both based on the validation loss or validation ROUGE score). For each of the models, we performed hyperparameter optimization on the learning rate. We started 10 training runs per model with varying learning rates (1e-3 to 1e-6). We trained each model on a single RTX A6000 GPU with an effective batch size of 8. We use the Adam optimizer (Kingma and Ba, 2015) and warm up the learning rate for 2 epochs. As our models can only process inputs of 1024 tokens, we filter the announcements with an oracle that selects the sentences that maximize the ROUGE-2 score when compared to the annotation (Liu and Lapata, 2019). Because of the lack of available pretrained long-context models, we leave ingesting the entire announcement into the model for future work.

4.3 Inference Details

When generating with the sequence-to-sequence models, we use beam search with 5 beams. We generate at least 100 and at most 512 tokens. We use n-gram blocking with $n = 6$ (Paulus et al., 2018).

For LLMs, we provide 5 in-context examples of an announcement with a corresponding interpretation, taken from the training set. We then prompt with the evaluation announcement. The announcements are prefixed with *Announcement:*, and the interpretations with *Interpretation:*. We generate with nucleus (top_p) sampling (Holtzman et al., 2020). We vary the temperature (0.5, 0.7, 1.0, 2.0) and the top_p (0.5, 0.8, 0.9, 0.95). We compare loading the model’s weights in float32 with bfloat16 (16-bit precision shows a small performance drop). For instruction-tuned LLaMA, we also try using a system prompt³ (not available for

³You are a chatbot that analyzes political announcements and replies with a coded interpretation of its main points."

Gemma). In total, we try 12 hyperparameter combinations per LLM. The best settings can be found in Appendix B.

5 Results

The results in Table 2 show that the BART (large) model performs best among the models we tried. It scores the highest on lexical similarity (ROUGE) and semantic similarity (BERTScore), and achieves reasonable diversity and novelty of generated text. Appendix C shows an example output. We now present our main findings from these results.

Increasing model size does not improve results.

We experimented with different-sized sequence-to-sequence models. We cannot make out a general trend in the change of performance due to model size. The best-performing model BART is of medium size. Additionally, LLMs do not reach the performance of the finetuned smaller models on this task.

Instruction tuning helps in-context learning.

For both LLaMA and Gemma, the instruction-tuned versions massively outperform the base models on all metrics that measure similarity with the reference interpretations. The negative rescaled scores of the base LLMs suggest that content-wise, the LLMs’ generations are less semantically similar to the ground truth than two randomly drawn sentences from Common Crawl (cf. Zhang et al., 2020). It seems that instruction tuning is a necessary ingredient of LLM training to enable in-context learning on this task.

Instruction tuning drastically shortens outputs.

While not listed in Table 2, we also find that instruction-tuned models generate shorter outputs (less than half the tokens than the base models, and even shorter than the sequence-to-sequence

Model	Std sent	Act	Motive	Evidence	Russia link	Correct format
References	1.55 (\pm 0.85)	1.57 (\pm 0.89)	0.21 (\pm 0.57)	0.60 (\pm 0.93)	1.58 (\pm 0.92)	99.41%
T5 (base)	1.54 (\pm 0.66)	0.62 (\pm 0.53)	0.00 (\pm 0.00)	0.00 (\pm 0.00)	0.15 (\pm 0.37)	85.67%
Flan-T5 (base)	1.48 (\pm 0.59)	1.11 (\pm 0.67)	0.00 (\pm 0.00)	0.05 (\pm 0.25)	0.37 (\pm 0.56)	94.56%
BART (large)	1.78 (\pm 0.59)	1.54 (\pm 0.60)	0.06 (\pm 0.26)	0.56 (\pm 0.89)	1.29 (\pm 0.60)	96.68%
T5 (large)	1.42 (\pm 0.54)	0.47 (\pm 0.52)	0.00 (\pm 0.00)	0.00 (\pm 0.00)	0.30 (\pm 0.50)	85.78%
Flan-T5 (large)	1.70 (\pm 0.61)	1.23 (\pm 0.64)	0.00 (\pm 0.06)	0.06 (\pm 0.25)	0.41 (\pm 0.58)	89.58%
In-context examples	1.2	1.2	0	0	1.2	
LLaMA-3.1	0.13 (\pm 0.33)	0.21 (\pm 0.47)	0.00 (\pm 0.00)	0.12 (\pm 0.39)	0.16 (\pm 0.39)	61.53%
LLaMA-3.1 (instruct)	1.03 (\pm 0.70)	0.70 (\pm 0.62)	0.00 (\pm 0.00)	0.39 (\pm 0.69)	0.61 (\pm 0.60)	89.60%
Gemma-2	0.23 (\pm 0.47)	0.17 (\pm 0.39)	0.00 (\pm 0.00)	0.08 (\pm 0.30)	0.13 (\pm 0.33)	78.72%
Gemma-2 (instruct)	1.17 (\pm 0.68)	0.59 (\pm 0.51)	0.00 (\pm 0.00)	0.07 (\pm 0.27)	0.63 (\pm 0.51)	95.55%

Table 3: Selected annotation counts with standard deviation on the test set.

models). They also produce many more annotation start and end marker tokens (around 10% of total tokens), whereas base models generate only around 2%. References contain 12.4% of these special tokens.

Table 3 shows the counts and standard deviation for a selection of annotation categories. Again, the BART model matches the reference distribution the closest (except for standardized sentences). Motives are underrepresented in all model outputs. In the last column, we also report if models correctly open and close annotations with matching start and end markers⁴.

Only BART follows the reference category distribution. All sequence-to-sequence models and the instruction-tuned LLMs generate more than one standardized sentence on average. However, except for BART, models seldom generate all the mandatory categories of an interpretation. Moreover, BART generates the correct format more often than any other model.

Instruction tuning is vital for learning the format and distribution. For both LLaMA and Gemma, there is a major difference between the base model and the instruction-tuned model. Instruction tuning both allows the models to pick up the distribution of annotation categories and the annotation format with start and end markers much better, reaching similar performance as the sequence-to-sequence models. For some categories (evidence, Russia link), they generate more annotations than the T5 and Flan-T5 models. All this is achieved with only 5 in-context examples.

⁴References are not 100% correct since if they are too long (we used 512 tokens), the matching end markers get cut off.

LLMs generate categories that are not in the in-context examples. Surprisingly, we found that LLMs also generated categories not present in our in-context examples. Especially evidence is generated quite frequently, particularly by LLaMA (instruct). A natural explanation is that the LLMs must have been pretrained on a similar dataset that contained those or similar annotations. If this is the case, it is still interesting to see that both LLMs transfer that pretraining knowledge so readily. The annotation category scope, which is not shown in Table 3, appears twice in the in-context examples, yet is generated fewer times by all LLMs except the base LLaMA. This, however, also means that even though our dataset has not been released yet, performance on it will depend on whether the used models have been pretrained on similar datasets.

On top of the annotation categories present in our dataset, LLaMA also generates the additional categories *location* and *source*, while Gemma generates *location* and *organization*. They are, however, very rare, appearing at most 4 times for our total of 264 announcements in the test set.

6 Conclusion

We presented a new dataset on generating interpretations for policy announcements concerning US foreign policy with respect to Russia for the years 1993 to 2016. We evaluated common language models on this generation task and found that fine-tuned sequence-to-sequence models, specifically BART, outperformed few-shot prompted large language models.

Interesting directions for future work are investigating long-context methods that can access the entire announcement when writing the interpretation, and a more structured approach to generating the individual categories and their contents. We also

hope to see whether models trained on this dataset can be successfully transferred to other tasks and domains.

Limitations

We discuss limitations of our work in the following.

Limitations of the dataset. While the target interpretations are carefully curated by human experts, the source announcements are extracted from PDFs and websites. Especially text extracted from the former may include artifacts, such as additional or missing whitespace and punctuation caused by unusual formatting, or only partially extracted text.

Limitations of evaluated methods. We believe that the performance of LLMs could be improved in various ways. Since the presented task is rather complex, more few-shot examples could be given as additional context for the model to adapt. In some cases, Gemma even asked for more information in its response, e.g. "I am still under development and learning to interpret complex text. Can you please provide me with more context or specify what you would like me to do with this text?" Furthermore, parameter-efficient finetuning of LLMs (e.g. low-rank adaptation; [Hu et al., 2022](#)) may further improve their results. We leave these investigations to future work.

Limitations of evaluation metrics. In this paper, we evaluate models with established automatic metrics for text generation. For the task of generating interpretations, metrics comparing a candidate to a reference interpretation by lexical or semantic similarity will naturally miss the more subtle aspects. An evaluation that extracts the main characteristics of the interpretation in a broader context is interesting for future work.

Acknowledgments

This work was supported as a part of the grant Automated interpretation of political and economic policy documents: Machine learning using semantic and syntactic information, funded by the Swiss National Science Foundation (grant number CRSII5_180320), and led by the co-PIs James Henderson, Jean-Louis Arcand and David Sylvan.

References

Gavin Abercrombie and Riza Batista-Navarro. 2020. [ParlVote: A corpus for sentiment analysis of po-](#)

[litical debates](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. [It’s not rocket science: Interpreting figurative language in narratives](#). *Transactions of the Association for Computational Linguistics*, 10:589–606.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.

Taeyoung Doh, Sungil Kim, and Shu-Kuei Yang. 2021. [How you say it matters: Text analysis of fomc statements using natural language processing](#). *Economic Review-Federal Reserve Bank of Kansas City*, 106(1):25–40.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.

Stefan Fuchs, Michael Witbrock, Johannes Dimyadi, and Robert Amor. 2023. Using large language models for the interpretation of building regulations. In *13th Conference on Engineering, Project and Production Management*.

Max Goplerud. 2021. [Methods for analyzing parliamentary debates](#). In *The Politics of Legislative Debates*, chapter 5. Oxford University Press.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- David A. Hoffman and Yonathan A. Arbel. 2023. [Generative interpretation](#). *New York University Law Review*, 99.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Jean Lee, Hoyoul Luis Youn, Nicholas Stevens, Josiah Poon, and Soyeon Caren Han. 2021. [Fednlp: An interpretable nlp system to decode federal reserve communications](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2560–2564, New York, NY, USA. Association for Computing Machinery.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. [Testing the ability of language models to interpret figurative language](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.
- Yang Liu and Mirella Lapata. 2019. [Text summarization with pretrained encoders](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.
- Andreas Marfurt, Ashley Thornton, David Sylvan, Lonneke van der Plas, and James Henderson. 2022. [A corpus and evaluation for predicting semi-structured human annotations](#). In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 262–275, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2018. [A deep reinforced model for abstractive summarization](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. [How to do politics with words: Investigating speech acts in parliamentary debates](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. [Galactica: A large language model for science](#). *Preprint*, arXiv:2211.09085.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Brydon T Wang. 2024. [Prompts and large language models: A new tool for drafting, reviewing and interpreting contracts?](#) *Law, Technology and Humans*, 6(2):88–106.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Dataset Creation

We outline the details of dataset creation. First, the New York Times archives were searched for articles on the foreign policy of the US with respect to Russia. In these articles, domain experts searched

Model	Early stopping	Max LR model	Max LR LM head
T5 (base)	Loss	1e-4	1e-3
Flan-T5 (base)	ROUGE	1e-4	1e-4
BART (large)	ROUGE	1e-5	1e-4
T5 (large)	ROUGE	1e-4	1e-4
Flan-T5 (large)	ROUGE	1e-4	1e-3

Table 4: Best hyperparameter settings for sequence-to-sequence models based on the validation set.

Model	Temperature	Top_p
LLaMA-3.1	1.0	0.9
LLaMA-3.1 (instruct)	0.7	0.5
Gemma-2	0.7	0.95
Gemma-2 (instruct)	0.7	0.9

Table 5: Best hyperparameter settings for LLMs based on the validation set.

for partial, complete or multiple sentences that contain all the information for the required annotation categories described in Section 3. All categories were then marked and potentially commented on to surface the readers’ necessary background knowledge. Then, they are validated by a senior domain expert.

B Hyperparameter Settings

We list the optimal hyperparameter settings for each of our models in Tables 4 and 5.

C Example Outputs

In Table 6, we show a source announcement and the corresponding reference interpretation and interpretations of BART and Gemma-2 (instruct). In this example, BART focuses on a different part of the speech than the reference interpretation, which could have nevertheless been picked up in an article. Gemma gets the main point right but becomes too repetitive. BART uses annotation categories in the right places but does not close them correctly in the second sentence. Gemma closes all annotations correctly but places them on parts of the text that do not match.

Source announcement

From State Department Dispatch, Vol. 4, No. 3, 1993: Chemical Weapons Convention Signing Ceremony Secretary Eagleburger Remarks upon signing the Chemical Weapons Convention, Paris, France, January 13, 1993.

It is fitting that we meet to sign this historic Chemical Weapons Convention in a city where, 4 years ago, the international community appealed for the strengthening of norms against chemical warfare. I am pleased to be in Paris, and I am especially pleased to represent my President, George Bush, a man who, over the course of the past decade, launched some of the key initiatives which helped to make this agreement possible. He and all those responsible can take pride in an achievement whose revolutionary scope and impact we can recognize today without having to await the verdict of history. But such has been the amazing record of the past few years. We have seen the international community liberate itself from half a century of gridlock and paralysis and move beyond the rhetoric of democracy to achieve real democracy; move beyond the rhetoric of detente to achieve real peace; and move beyond the rhetoric of disarmament to achieve real reductions in weapons of mass destruction. The Chemical Weapons Convention we sign today does more than simply reduce a class of arms or mitigate against their proliferation. This convention mandates a worldwide non-discriminatory ban on an entire class of weapons of mass destruction—the only class of such weapons that has been widely used in combat. By the radical terms of this agreement, all signatory states forswear the possession, production, stockpiling, transfer, and, indeed, the use of chemical weapons; and all signatories must destroy all chemical weapons and chemical weapons production facilities in their possession. Moreover, the convention's strict verification regime, which accommodates legitimate commercial and sovereign interests, sets an innovative standard for future multilateral agreements. The international community is virtually united in support of the objectives of the Chemical Weapons Convention. However, there must be truly global adherence if the convention is to achieve its purpose and if doubts are to be eliminated over the commitment and intentions of those who fail to sign, ratify, and fully comply with its terms. Nowhere is this more important today than in the Middle East, a region which over the past 30 years has been home to more active chemical weapons programs—and which has seen more chemical weapons use—than any other part of the world. It is, therefore, particularly disappointing that so many Middle Eastern states are absent from this ceremony today. The fact of the matter is that linking this convention to other issues cannot affect the fate of those issues, but it will surely undermine the effect of this treaty in the one region most exposed to the danger of chemical weapons—namely, the Middle East. The point, I believe, is to tackle the challenge of weapons of mass destruction wherever we can, whenever we can. I would, therefore, urge the members of the Arab League to seize this opportunity and sign the Chemical Weapons Convention. Doing so would be a step toward, and not away from, making the Middle East a zone free of all weapons of mass destruction, as called for by President Mubarak of Egypt. Today's ceremony is only the beginning of the work which lies ahead. Next month, the Preparatory Commission will meet in The Hague [the Netherlands] to work out the important and detailed provisions for implementing the convention. The United States is fully committed to the success of those efforts, which will require the same broad support and participation which produced the successful convention itself. As I indicated at the beginning, the past few years have been a remarkably creative period of international achievement. Perhaps not coincidentally, I believe that President Bush's passage across the international scene has equally been one of tangible achievement, particularly in terms of the issue most important to the fate and future of the planet—the issue of weapons of mass destruction. George Bush's legacy will include landmark treaties—START [Strategic Arms Reduction Treaty] I, START II, and CFE [Conventional Armed Forces in Europe]—as well as diplomatic efforts which paid non-proliferation dividends in Africa, South America, the Middle East, and here in Paris today. But he knows, as all of us must know, that what we have accomplished to date will matter little unless we are prepared to confront the even greater proliferation dangers we most certainly will face in the years to come.

Reference interpretation

[STD SENTENCE START] [ACT START] [RUSSIA LINK START] [ACTOR START] The United States (USA) [ACTOR END] and Russia were among those signing [REFERENCE START] today (The US signed the Chemical Weapons Convention) [ACT END] (Russia also signed the Chemical Weapons Convention) [RUSSIA LINK END] [REFERENCE END] . Speaking on behalf of the United States, which has already begun destroying its chemical-weapons stockpiles, Secretary of State Lawrence S. Eagleburger noted that [EVIDENCE START] the treaty mandates a global ban on the only class of weapons of mass destruction "that has been widely used in combat." [EVIDENCE END] [STD SENTENCE END]
[STD SENTENCE START] In [EVIDENCE START] what has been hailed as the first attempt to eliminate an entire category of weapons of mass destruction [EVIDENCE END] , [ACT START] [RUSSIA LINK START] more than 120 nations, including [ACTOR START] the United States (USA) [ACTOR END] and Russia, [REFERENCE START] began today [REFERENCE END] to sign an agreement to ban the production, stockpiling and use of chemical weapons (The US signed the Chemical Weapons Convention) [ACT END] (Russia also signed the Chemical Weapons Convention) [RUSSIA LINK END] . [STD SENTENCE END]

Candidate interpretation by BART

[STD SENTENCE START] [ACTOR START] Secretary of State Warren Christopher (United States) [ACTOR END] [ACT START] [REFERENCE START] urged [REFERENCE END] the Arab League today to join the [RUSSIA LINK START] Chemical Weapons Convention (Russia is a signatory of the Chemical Weapons Convention) [RUSSIA LINK END], which bans chemical weapons (The US urged the Arab League to sign the Chemical Weapons Convention and thus become part of the WMD regime) [ACT END]. [STD SENTENCE END]
[STD SENTENCE START] [ACTOR START] Mr. Christopher (United States [ACTOR END] [REFERENCE START] said [REFERENCE END], [ACT START] " [EVIDENCE START] The point, I believe, is to tackle the challenge of weapons of mass destruction wherever we can, whenever we can [EVIDENCE END]. Nowhere is this more important today than in the Middle East, a region which over the past 30 years has been home to more active chemical weapons programs – and which has seen more chemical weapons use – than any other part of the world [EVIDENCE END]." [STD SENTENCE END]
ROUGE-full (1/2/L): 55.22/20.34/41.75, BERTScore (rescaled): 15.20, novel 2-grams: 46.43%

Candidate interpretation by Gemma-2 (instruct)

[STD SENTENCE START] [RUSSIA LINK START] Mr. Eagleburger (USA) [RUSSIA LINK END] stressed that [ACT START] [REFERENCE START] the United States [REFERENCE END] [ACTOR START] is committed to [ACTOR END] working with all nations to achieve the universal ratification and implementation of the Chemical Weapons Convention (The United States is committed to working with all nations to achieve the universal ratification and implementation of the Chemical Weapons Convention) [ACT END] (The United States is committed to working with all nations to achieve the universal ratification and implementation of the Chemical Weapons Convention) [STD SENTENCE END] .
ROUGE-full (1/2/L): 39.33/14.35/30.13, BERTScore (rescaled): 12.65, novel 2-grams: 74.19%

Table 6: Example outputs by the best finetuned sequence-to-sequence model (BART) and the best few-shot prompted LLM (Gemma-2) on the first example of the validation set, together with their evaluation scores on this specific example.

Order Up! Micromanaging Inconsistencies in ChatGPT-4o Text Analyses

Erkki Mervaala^{1,2} and Ilona Kousa¹

¹ University of Helsinki, Finland

² Finnish Environment Institute Syke, Finland

firstname.lastname@helsinki.fi

Abstract

Large language model (LLM) applications have taken the world by storm in the past two years, and the academic sphere has not been an exception. One common, cumbersome task for researchers to attempt to automatise has been text annotation and, to an extent, analysis. Popular LLMs such as ChatGPT have been examined as a research assistant and as an analysis tool, and several discrepancies regarding both transparency and the generative content have been uncovered. Our research approaches the usability and trustworthiness of ChatGPT for text analysis, more specifically keyword extraction and topic classification, from the point of view of an "out-of-the-box" zero-shot or few-shot setting, focusing on how the context window and mixed text types affect the analyses generated. Results from our testing indicate that both the types of the texts and the ordering of different kinds of texts do affect the ChatGPT analysis, but also that the context-building is less likely to cause analysis deterioration when analysing similar texts. Though some of these issues are at the core of how LLMs function, many of these caveats can be addressed by transparent research planning.

1 Introduction

The potential for utilising large language models for data analysis has generated widespread interest among researchers across various fields (Azaria et al., 2024; Hadi et al., 2023), and one of the most extensively studied is the field's market-leader OpenAI's ChatGPT

(Ray, 2023; Westfall, 2023). As a freely accessible tool with a simple user interface, it has brought a new repertoire of analytical methods within reach of researchers who may have previously faced technological or resource barriers to using computational social science methods.

Several researchers have examined ChatGPT's capabilities as a data analysis tool and reported their findings, striving to approach the method critically from the perspective of scientific principles and share best practices with other researchers (Bilal et al., 2024; Törnberg, 2023). Despite the hype surrounding the tool, ChatGPT's, and large language models in general, have been criticised and found to be unreliable in various research tasks including text analysis and annotation (Ollion et al., 2023). Our paper is an additional contribution to this literature from the point of view of automated text analysis, but also an addition to the yet very limited research focusing on how such services function with smaller languages (Mets et al., 2024).

The research aims to evaluate whether the keyword extraction and topic classification produced by chat-based large language model (LLM) services, such as ChatGPT, vary based on different content types, the mixture of contents, the order of mixed contents, and context building within the same conversation window.

71 Our purpose is to gain an understanding of
72 the influence of context on classification and
73 to shed light on what kind of biases and
74 limitations must be considered in research
75 employing LLM tools, especially when using
76 several different types of data.

77 The specific case under examination will
78 focus on the discussion regarding the green
79 transition, which in political rhetoric generally
80 refers to a shift towards an ecologically
81 sustainable and low-carbon economy
82 (Filipović et al., 2022; Ministry of
83 Environment Finland, 2024). Green transition
84 is an interesting term in the political debate as
85 it evokes very polarising reactions from highly
86 supportive and positive to strongly negative
87 and dismissive. The reactions also vary a lot by
88 country: for example, in Norway, the term
89 “green transition” has very positive
90 connotations and was even voted as the word
91 of the year in 2015 (Olerud et al., 2016).

92 In Finland, the government in power since
93 2023 led by the National Coalition Party and
94 its support party The Finns has made efforts to
95 replace the term with other expressions such as
96 "clean transition" (“puhdas siirtymä”) or “the
97 blue-white transition” (“sinivalkoinen
98 siirtymä”) (Tavio, 2023; Valtioneuvosto,
99 2023).

100 To study the debate from the perspectives of
101 both politicians and the public, we analysed
102 communication on Twitter and within the
103 Finnish parliament during the Conference of
104 Parties climate change conference (COP27)
105 held in Egypt in late 2022. We chose the highly
106 publicised international event as green
107 transition was one of its core focus points
108 (European Commission, 2022) and because it
109 stirred conversation both online and in the
110 parliament so close to the Finnish
111 parliamentary elections of 2023. The data
112 comprises parliamentary speeches and Twitter
113 comments collected from October 19, 2022, to
114 December 1, 2022. Each dataset contains 20
115 texts, resulting in a total of 20 tweets (T) and
116 20 parliamentary speeches (P).

117 The study expects to determine if and how
118 the order and context in which texts are

119 presented to the LLM influence the outcomes
120 of text analysis, specifically in terms of
121 keyword extraction and topic classification.
122 Our hypotheses address the usability of LLMs
123 for our specific text analysis task from three
124 different angles:

125 Hypothesis 1: It is possible to gain
126 meaningful results using LLMs in analysing
127 and identifying keywords and topics in textual
128 data in a zero-shot or a few-shot setting,
129 focusing on a specific political issue such as
130 “green transition”.

131 Hypothesis 2: The initial content type will
132 influence the keyword extraction and topic
133 classification when texts are analysed in a
134 mixed order.

135 Hypothesis 3: The context-building affects
136 the results when analysing texts in a few-shot
137 setting within the same chat window compared
138 to zero-shot analyses in separate chat
139 windows.

140 1.1 Review of previous research

141 The consistency of responses generated by
142 ChatGPT have been empirically studied in
143 contexts such as medicine (Lechien et al.,
144 2024), mathematics (Heya et al., 2024) and
145 coding (Clark et al., 2024).

146 In addition, ChatGPT's model is known to
147 consider the order of the input tokens when
148 generating its responses (Bansal et al., 2024),
149 which has been empirically demonstrated, for
150 example, by varying the word order of the
151 input (Zhao et al., 2024). The effects of order-
152 dependency on consistency of the output have
153 been studied further for example by Jang and
154 Lukasiewicz, who showed that ChatGPT and
155 GPT-4 models often failed in both semantic
156 and symmetric consistency, meaning that they
157 produced different results from semantically
158 similar inputs, and that the sentence order of
159 the input affected the predictions made by the
160 models.

161 For text analysis and annotation specifically,
162 it has been shown that zero-shot or few-shot
163 approaches to utilising LLMs often fail to
164 reach as accurate results as fine-tuned, human-
165 annotated models (Ollion et al., 2023). Other

166 serious issues raised in the literature include
167 the unreliability, potentially “hallucinated”
168 results, copyright issues, and stochastic
169 generation of misinformation and false claims
170 (Alkaissi and McFarlane, 2023; Guerreiro et
171 al., 2022; Guerreiro et al., 2023; Khatun and
172 Brown, 2023).

173 2 Methods and materials

174 The parliamentary speeches were obtained via
175 the Finnish parliamentary speech archive
176 ParlamentiSampo (Hyvönen et al., 2022) that
177 allows collecting speeches from a selected
178 time period in .csv format and then filtering the
179 dataset via declension of the key phrase
180 “vihreä siirtymä” (“green transition”). The
181 filtering left us with 20 parliamentary speeches
182 that fit the criteria.

183 The tweets for the study had been collected
184 via the now-defunct service Mohawk Analysis.
185 As there were several thousands of tweets
186 containing the key phrase, the tweets were
187 chosen first by matching the dates of the
188 parliamentary speeches. The number of tweets
189 for the studied time period was much larger
190 than parliamentary speeches, so a
191 corresponding tweet was picked via a
192 randomised process for each speech. For
193 example, when there were five speeches on
194 October 27th, 2022, a total of five tweets were
195 selected from the Twitter dataset. The only
196 limiting factor for a tweet was that it should be
197 a standalone tweet and not a response to
198 another user. Links to both parliamentary
199 speeches and the tweets in question are found
200 in Appendix A.

201 For the purposes of replicability, each
202 analysis was done via the native commercial
203 and publicly freely available ChatGPT
204 service’s version ChatGPT-4o released in May
205 2024. Its context window is 128,000 tokens,
206 which was enough for all our test setups
207 (OpenAI, 2024b). The initial setup settings for
208 the model were attained in June via prompting
209 (see Appendix B).

210 All tests except 1.3 were conducted between
211 July 8 and August 4, 2024, by one author, and

212 initial tests of the analysis environment were
213 performed between June 10 and July 21, 2024.
214 Originally, the environment initialisation
215 prompt was performed to control the changes
216 in the ChatGPT version (see Appendix B) but
217 access to it by users was later removed. The
218 other author completed test 1.3 between
219 August 19 and 26, 2024.

220 Each analysis began with the same first
221 prompt after which the concurrent prompts
222 included only the text to be analysed. No
223 feedback was given to ChatGPT during the
224 tests. If after prompting ChatGPT offered
225 options to choose our preferred answer, this
226 would be ignored, and the next text was added
227 to the chat instead. The first prompt including
228 the instructions for text analysis is as follows:

229 *You will be analyzing a series of texts. For*
230 *each text, perform the following analysis: 1.*
231 *Extract a list of keywords from the text. 2. For*
232 *each extracted keyword, provide a*
233 *corresponding topic category.*

234 2.1 Test batch 1: zero-shot and few-shot 235 analyses

236 This experimental scenario evaluated how the
237 text analysis differs when all content is
238 analysed in the same chat window in a few-
239 shot setting versus each item being analysed in
240 separate chat windows in a zero-shot setting.

241 In the context of ChatGPT, a zero-shot
242 setting refers to the language model’s ability to
243 perform a task or respond to a query without
244 having been explicitly trained on examples of
245 that specific task or scenario. Instead, the
246 model relies on its general understanding of
247 language and knowledge to generate an
248 appropriate response based on the prompt
249 alone. (Yuan et al., 2023)

250 In test 1.1 all parliamentary speeches were
251 analysed in the same chat window, and in test
252 1.2 all tweets were analysed in the same chat
253 window, allowing for the window-specific
254 context to build throughout both analyses runs
255 covering each set of 20 texts. Tests 1.3 and 1.4
256 included the texts of both datasets analysed in
257 separate chat windows: test 1.3 included 20
258 prompts in 20 chat windows, one for each

259 parliamentary speech, and test 1.4 had a 306
260 similar setup for tweets.

261 The rationale behind the testing addresses 308
262 the issue with a zero-shot setting where the 309
263 LLM is given the task of text analysis without 310
264 context from earlier text analysis tasks. In the 311
265 one-window task, all the texts were submitted 312
266 for analysis consecutively which builds
267 context cumulatively, thus potentially 313
268 enhancing the analysis the further the
269 conversation goes. The hypothesis is that in
270 analyses performed within the same chat
271 window the first and the last text are,
272 essentially, analysed by a different LLM as the
273 context has been allowed to build. To test this,
274 the texts were presented in the chat first in the
275 original, chronological order, and then in a
276 backwards order. It is due to this feature that
277 leads to the conclusion that the only strictly
278 zero-shot analyses to be made via ChatGPT are
279 to be done individually in separate chat
280 windows. The individual analyses were then
281 compared to the analyses done within the same
282 chat window.

283 It should also be pointed out that the cross-
284 chat “Memory” feature for Plus tier ChatGPT
285 subscribers was not available in the EU area
286 during the testing and so had no impact on it.
287 The feature that allows ChatGPT to
288 “remember details between chats, allowing it
289 to provide more relevant responses” has was
290 made available in EU in September 2024 327
291 (Coombes 2024, OpenAI 2024c).

292 2.2 Test batch 2: mixed datasets

293 This second experiment evaluated how the
294 context building within the same chat window
295 affects text analysis results when all 40 texts
296 are analysed in a mixed order, and whether the
297 initial content type influences the outcome.

298 Test 2.1 included all tweets and
299 parliamentary speeches alternating, starting
300 from a tweet. Test 2.2 was almost identical but
301 this time the first input was a parliamentary
302 speech. Test 2.3 began with the whole tweet
303 dataset, which was then followed by the whole
304 speech dataset, and test 2.4 began with the
305 speeches and ended with the tweets.

306 The rationale behind the order-setting pairs
307 stems from the context-building nature of the
308 LLM chat windows. In the tests, the context is
309 first built with consecutive texts of the same
310 type and then the type changes to a very
311 different one - both in tone and in length.

313 3 Results

314 In this section, we first describe the
315 quantitative results of keyword extraction and
316 topic classification and then evaluate the
317 consistency and relevance classification.

318 3.1 Keywords and topics

319 Overall, the amounts of keywords and topics
320 identified by ChatGPT varied. In the first batch
321 of tests with separated datasets (see Table 1),
322 both analyses of parliamentary speeches (PS)
323 provided more keywords and topics than
324 tweets (T), which is understandable due to the
325 speeches being much longer.

Test #	Keywords	Topics
1.1 (PS, few-shot)	439	414
1.2 (T, few-shot)	127	125
1.3 (PS, zero-shot)	376	348
1.4 (T, zero-shot)	135	121

Table 1: Keywords and topics in test batch 1.

327 In our analysis, less keywords and topics
328 were identified in the zero-shot analysis (zs)
329 than the few-shot analysis (fs), but the zero-
330 shot tweet analysis found more keywords but
331 less topics than the few-shot analysis.

332 Test batch 2 with mixed datasets (Table 2)
333 expectedly produced more keywords and
334 topics, as each analysis had double the number
335 of texts. The largest amounts of keywords and
336 topics were found in test 2.3 that analysed first
337 all 20 tweets and then 20 speeches (T1–T20,
338 PS1–PS20), followed by test 2.1 which
339 alternated between content type but started
340 with a tweet (T1, PS1...PS20).

Test #	Key-words	Topics
2.1 (T1,PS1...PS20)	619	600
2.2 (PS1,T1...T20)	449	310
2.3 (T1-T20,PS1-PS20)	708	643
2.4 (PS1-PS20,T1-T20)	536	338

Table 2: Keywords and topics in test batch 2.

Despite the same content analysed within the same content window, the order of the texts did impact the amounts of keywords and topics found. In these specific cases, tests starting with a tweet produced over a hundred more keywords and topics identified by ChatGPT.

3.2 Classification of “green transition”

In the consistency and relevance evaluation, we compared the classification results of tests 1.1 to 1.4 (see Table 3). We studied how consistently the keyword "green transition" was classified in Twitter and Parliamentary speech datasets, in a few-shot versus a zero-shot setting. There was some variation in whether "green transition" was recognized as a keyword at all or in a slightly different form (e.g. "left-green transition"), which is why the total values vary between 17 and 19.

Topic	1.1 (PS, few-shot)	1.2 (T, few-shot)	1.3 (PS, zero-shot)	1.4 (T, zero-shot)
Environmental Policy	8	18	4	6
Green Transition	9	0	2	0
Environmental Policy/ Sustainability	0	0	0	4
Environmental Policy, Sustainable Development	0	0	2	0
Other	0	0	11	10
Total	17	18	19	18

Table 3: Topics for the keyword “green transition” in test batch 1.

For both tweets and parliamentary speeches, classification was more consistent when all texts were analysed in a few-shot setting in the same chat window: all Tweets were classified into the category Environmental Policy and parliamentary speeches were classified into two different categories: Environmental Policy and Green Transition. In the zero-shot setting, there were 12 different category names for the keyword “green transition” for tweets and 11 different category names for parliamentary speeches. Another observation was that few-shot analyses did not contain any multi-level categories (such as “Environmental Policy / Sustainability”), whereas in zero-shot setting, multi-level categories appeared in 12 cases for tweets and 11 cases for parliamentary speeches.

The results indicate that the topic categories were quite sensitive to variation in the form of the keyword. For example, in the single-window parliamentary speeches test, the keyword "green transition" was consistently categorised into Environmental Politics, until P11, where the keyword and topic category were in a slightly different form: Green Digital Transition. In subsequent analyses, the topic category for “green transition” changed from Environmental Politics to Green Transition.

The results confirm our first hypothesis: the classification of “green transition” was relevant and meaningful in these cases. Most of the time, the name of the category was very general (e.g. “Environmental Policy”), but sometimes more specific (e.g. “Energy Transition”). The usability and preferred specificity of the categories depends on the context, but in any case, all categories corresponding to “green transition” were named in a relevant way. However, the results suggest that the classification may be influenced by the specific phrasing of the keywords used in the analysis, which can significantly affect the results especially in languages like Finnish, where compound words are common.

410 3.3 Examples of context-building and errors

411 Test 1.1 that had all parliamentary speeches
412 analysed within the same chat interaction
413 showed no deterioration during the
414 conversation. The results were consistent and
415 delivered the more keywords and categories
416 the longer the speeches were. Test 1.3 provided
417 similar results, but the topic categories were
418 more often multi-leveled such as “Economy /
419 Finance”. Test 1.4 with separate tweet analyses
420 showed slight variation in style and included
421 multi-level categories but no significant
422 mistakes or errors related to the task.

423 Test 1.2 however had some significant
424 issues and differences compared to how the
425 context is built throughout test 1.1 and how the
426 individual tweet analyses of test 1.4. As the
427 task was going through all the 20 tweets, the
428 beginning seemed consistent until T7 that
429 began with an English word “sorry” that led
430 ChatGPT to indeed answer the comment in a
431 conversational manner in English rather than
432 analysing the text:

433 **T7:**

434 *"Sorry, kun jankutan, mutta vihreä*
435 *siirtymä pitää sisällään myös #huontokato*
436 *hillinnän ja luonnon vahvistamisen*
437 *#ilmastokriisi hillinnän rinnalla. Toista ei*
438 *voi ratkaista ilman, että ratkaistaan*
439 *molemmat. Mukavaa päivää!"*

440 **ChatGPT answer #1:**

441 *"No worries, your points are important!*
442 *The green transition indeed encompasses*
443 *both biodiversity conservation and climate*
444 *crisis mitigation. Think of it as a holistic*
445 *approach where rewilding urban areas with*
446 *vertical forests or creating floating solar*
447 *farms on restored wetlands could tackle*
448 *both issues simultaneously. Innovative*
449 *projects like these can regenerate nature*
450 *while reducing emissions. Let's keep*
451 *pushing for creative solutions that address*
452 *multiple environmental challenges together.*
453 *Have a great day!"*

454 ChatGPT then offered a second alternative
455 (see example in Appendix C) consistent with
456 the earlier analyses, and, as there was no
457 reaction to this preference query and as the test
458 was continued as before, the following
459 analyses had no significant issues apart from
460 missing a word play “palvisota” (“ham war”)
461 in a tweet (T20), referring to pig farms and
462 ham and the Winter War - “Talvisota”.
463 Additionally, the full texts of the tweets were
464 included within each answer.

465 Out of the mixed type tests, test 2.1 that
466 alternated between tweets and parliamentary
467 speeches delivered very consistent results. As
468 with earlier single-type tests on tweets, the
469 topics were again multi-level, but this time the
470 classification stayed consistent throughout
471 labelling also parliamentary speeches with
472 similar multi-level topics.

473 Test 2.2 began also very consistently, though
474 this time the topics had only one level. Starting
475 with a parliamentary speech, the texts
476 alternated and were categorised seemingly
477 correctly until T10 in which the tweet author
478 asks about the “sustainable basis and logic
479 behind it” and claims how building a wind
480 power plant consumes more energy than the
481 plant would produce during its lifespan.
482 Instead of analysing the text, ChatGPT began
483 searching the web and then addressing the
484 issue by delivering a well-structured brief on
485 the topic, headlined “Sustainable Basis and
486 Logic Behind the Green Transition”. The
487 following parliamentary speech P11 returned
488 to form, but then T11 prompted a different
489 result: this time ChatGPT identified five
490 keywords and topics but then continued to
491 generate “Analysis and Commentary” and
492 then reiterating the “Sustainable Basis and
493 Logic Behind the Green Transition” with some
494 of the same links and wordings it used when
495 answering the T10 prompt. This analysis
496 ended with a new section “Conclusion”. P12
497 was the last to include the original style of
498 keyword and topic listing, as after T12
499 delivered the similar verbal analysis segments
500 from before, also the following speeches
501 included “Analysis and Commentary” and
502
503
504

505 “Conclusion”. This pattern remained until the
506 end of the test with the addition of a “Further
507 reading” segment that contained links to
508 external sources from T17 onwards.

509 The second pair of test batch 2 started with
510 analysing all the tweets first and then analysing
511 all the speeches. The analysis remained
512 consistent throughout the test providing clear
513 lists of keywords in Finnish and then topics in
514 English, though oftentimes the topic ended up
515 being just an English translation of the
516 identified Finnish keyword. From the point of
517 view of the consistency of keyword extraction
518 and topic classification, test 2.3 fairs equally as
519 well as 2.1.

520 In the Test 2.4, there were again notable
521 discrepancies. The initial phase of the test was
522 identical to test 1.1 meaning it included all
523 parliamentary speeches prompted for analysis
524 consecutively after which the same chat
525 window would be prompted to analyse all the
526 tweets. From the first tweet onwards, ChatGPT
527 began to change the way it answers. For the
528 first tweet, it comes up with 6 keywords and
529 their corresponding topic categories which is
530 significantly less than the 16–40 keywords and
531 categories of the previous parliamentary
532 speeches but, as previously mentioned,
533 understandable since the tweets are also
534 shorter.

535 However, ChatGPT included a new part of
536 analysis that intends to verbalise the analysis
537 of the content. The second tweet saw a
538 complete deterioration of the analysis process
539 as the tweet included a question whether
540 Finnish companies manufacture solar panels,
541 heat pumps and wind power (see Appendix C).
542 Instead of any keyword or topic category
543 extraction, ChatGPT began to answer the
544 question in Finnish by providing examples of
545 different technologies and details the export
546 technologies within the green transition for
547 Finland. The third tweet returned to form and
548 provided an analysis in a mixture of Finnish
549 and English with again more verbalised
550 analysis than before during the parliamentary
551 phase. This format remained with only slight

552 variations, ending each analysis with a
553 conclusion.

554 3.4 Other potential issues

555 Though there were no issues with submitting
556 long speeches for analysis in the chat window,
557 the seemingly random decision made by
558 ChatGPT of whether to include the original
559 text in the response or not affected directly
560 whether the full analysis of a given text
561 required additional prompting of the model to
562 “Continue generating” the full answer (see
563 appendix C). The longest speech was P10 with
564 835 words.

565 There were also occasional bugs in the chat
566 that prevented an analysis to complete or to run
567 at all, after which a new prompt was inserted.
568 Such bugs, and the issues with generation
569 limits, may also cause the prompt quota
570 reaching its limit which leads to interrupting
571 and pushing forwards finishing the current test
572 run.

573 Throughout all tests, one parliamentary
574 speech (P19) was flagged with a notice “This
575 content may violate our policies”. The speech
576 in question included terms such as “ihmisviha”
577 (“hatred of people”) and “Venäjä-viha”
578 (“hatred of Russia”). In both datasets, this one
579 single speech was the only one that was
580 flagged with such notification despite there
581 being other potentially “violating” content
582 present especially in the tweet dataset: one
583 tweet claimed that “green transition” was the
584 result of the “psychotic mind” of then prime
585 minister of Finland Sanna Marin (T18).

587 4 Conclusion

588 Based on the results of these experiments, our
589 initial hypotheses 2 and 3 regarding the context
590 window and its effects on text analysis proved
591 accurate. The order in which the different types
592 of texts were prompted for analysis influenced
593 the analyses happening in the same chat
594 window, and mixed datasets were also prone to
595 more errors compared to single-type analyses.
596 That being said, also the first hypothesis can be
597 said to be accurate as the tests, despite their

658 caveats, did provide meaningful insights about
659 the texts analysed.

600 As the current flagship product of the AI tool
601 market leader OpenAI (Westfall, 2023),
602 ChatGPT-4o can be said to represent the best-
603 performing, commercially available large
604 language models. Despite a relatively small
605 sample, our 240-prompt experiment shows
606 clearly that context-building is a significant
607 factor in text analysis performed with
608 ChatGPT-4o, adding to the list of varying,
609 valid reasons to question the usability of such
610 tools for text analysis in most cases. While the
611 individual “zero-shot” analyses do not risk
612 context deterioration, the “few-shot” may offer
613 more consistent classification especially if the
614 analysed texts are similar in type and length.
615 Additionally, the “zero-shot” analysis becomes
616 excessively cumbersome as the datasets grow.

617 For further research, and for the sake of
618 potential replicability, each such experiment as
619 ours should include transparent descriptions of
620 not only how each test was conducted but also
621 which version of the LLM was used and when,
622 especially when commercially available
623 services are used. It may not be possible to
624 return to the earlier versions of the GPT used
625 in the tests later. In fact, during the earlier
626 version of this paper, an analysis and
627 comparison between the legacy GPT-3.5 and
628 GPT-4 were planned and partially conducted.
629 OpenAI removed GPT-3.5 from the model
630 selection in July 2024 before all tests had been
631 completed (OpenAI, 2024d). This was not
632 accounted for as there was no generally
633 available information about the retirement date
634 for the model.

635 This relates to the larger issue of
636 replicability and transparency, and the
637 possibility for a stable research environment
638 which has not yet been possible to establish
639 using ChatGPT or similar kinds of chat-based
640 LLM services. Though the results themselves
641 cannot be trusted to be replicated due to the
642 very nature of large language models, even
643 with the exact same prompt, given the
644 potential for achieving very similar or even
645 identical analyses via accurate prompting and

646 low or no context building, the case for using
647 LLMs for text analysis can be made and
648 supported with certain caveats. A potentially
649 more accurate and replicable approach may
650 involve, for example, batch-processing
651 singular texts via application programming
652 interface (API) access, thus bypassing the
653 possibility of context building in the chat
654 window.

655 We also considered whether using a ready-
656 made classification framework created
657 manually or with another automated method,
658 such as topic modelling, would lead to more
659 consistent results. On the other hand, for
660 example, Bijker et al. (2024) found that data-
661 driven inductive classification worked more
662 reliably with ChatGPT compared to a ready-
663 made framework. However, the limitations of
664 the context window and the internal
665 inconsistency observed in this study could
666 hinder the development of a data-driven
667 classification framework. To tackle this
668 problem, consistency could possibly be
669 improved by first classifying a smaller subset
670 of data with ChatGPT, based on which a
671 framework could be created and used to
672 analyse a larger data set.

673 For chat-based systems, though, the changes
674 in the research environment are a large issue in
675 and of itself. Since November 2022, the public
676 version of ChatGPT has gone through many
677 changes (OpenAI, 2024a), and several of them
678 have not been transparently communicated to
679 its users (rafcin.s, 2024). Unreliable access to
680 the service or plugins - which may or may not
681 work at a given time - or other features of the
682 service, such as the periodically interrupted
683 access to the internet, and the abrupt changes
684 and updates to the available language models
685 have not allowed for a stable and sustainable
686 research environment to be developed.

687 Such approaches do still have to consider
688 the fact that the analysis or in fact any text
689 generated by an LLM cannot be ultimately
690 pass for a final product without human
691 assessment without risking potential
692 “hallucinations” or nonsensical “bullshit”

693 leaking into the text (Alkaissi and McFarlane,
694 2023; Hicks et al., 2024).

695 Overall, transparency both in how LLMs are
696 used and how LLMs work is seen as a high
697 priority, especially in academic and other
698 science-related work for which trustworthiness
699 is key (Nature, 2023; Ray, 2023). If the
700 previously mentioned caveats have been
701 considered, and the study authors have enough
702 expertise on how LLMs function and how to
703 ensure the research is both conducted
704 transparently, LLMs such as ChatGPT can be
705 powerful tools also for scientific research
706 (Azaria et al., 2024).

707

708 Acknowledgments

709 This study was supported by the Academy of
710 Finland, grant number 325207 (EM), the
711 Strategic Research Council, grant number
712 327768 (EM). Comments from the Puistokatu
713 4 and ORSI communities have been valuable
714 for the development of the analysis.

715

716 References

717 Hussam Alkaissi and Samy I McFarlane. 2023.
718 Artificial Hallucinations in ChatGPT: Implications
719 in Scientific Writing. *Cureus*.
720 <https://doi.org/10.7759/cureus.35179>

721 Amos Azaria, Rina Azoulay, and Shulamit Reches.
722 2024. ChatGPT is a Remarkable Tool—For Experts.
723 *Data Intelligence*, 6(1):240–296.
724 https://doi.org/10.1162/dint_a_00235

725 Gaurang Bansal, Vinay Chamola, Amir Hussain,
726 Mohsen Guizani, and Dusit Niyato. 2024.
727 Transforming Conversations with AI—A
728 Comprehensive Study of ChatGPT. *Cognitive
729 Computation*, 16(5):2487–2510.
730 <https://doi.org/10.1007/s12559-023-10236-2>

731 Iman Munire Bilal, Preslav Nakov, Rob Procter, and
732 Maria Liakata. 2024. Generating Zero-shot
733 Abstractive Explanations for Rumour Verification.
734 <https://doi.org/10.48550/ARXIV.2401.12713>

735 Autumn Clark, Daniel Igbokwe, Samantha Ross, and
736 Minhaz F. Zibran. 2024. A Quantitative Analysis of
737 Quality and Consistency in AI-generated Code. In
738 *2024 7th International Conference on Software and
739 System Engineering (ICoSSE)*, pages 37–41, Paris,
740 France. IEEE.
741 <https://doi.org/10.1109/ICoSSE62619.2024.00014>

742 European Commission. 2022. EU-Egypt Joint
743 Statement on Climate, Energy and Green Transition.
744 Available at:
745 [https://ec.europa.eu/commission/presscorner/detail/
746 en/statement_22_3703](https://ec.europa.eu/commission/presscorner/detail/en/statement_22_3703)

747 Lloyd Coombes. 2024. ChatGPT on Mac just got a big
748 update in the EU. Available at:
749 [https://www.msn.com/en-
750 us/news/technology/chatgpt-on-mac-just-got-a-big-
751 update-in-the-eu/ar-AA1qcIoa](https://www.msn.com/en-us/news/technology/chatgpt-on-mac-just-got-a-big-update-in-the-eu/ar-AA1qcIoa)

752 Sanja Filipović, Noam Lior, and Mirjana Radovanović.
753 2022. The green deal – just transition and
754 sustainable development goals Nexus. *Renewable
755 and Sustainable Energy Reviews*, 168:112759.
756 <https://doi.org/10.1016/j.rser.2022.112759>

757 Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf,
758 Barry Haddow, Alexandra Birch, Pierre Colombo,
759 and André F. T. Martins. 2023. Hallucinations in
760 Large Multilingual Translation Models.
761 <https://doi.org/10.48550/ARXIV.2303.16104>

762 Nuno M. Guerreiro, Elena Voita, and André F. T.
763 Martins. 2022. Looking for a Needle in a Haystack:
764 A Comprehensive Study of Hallucinations in Neural
765 Machine Translation.
766 <https://doi.org/10.48550/ARXIV.2208.05309>

767 Muhammad Usman Hadi, Qasem al Tashi, Rizwan
768 Qureshi, Abbas Shah, Amgad Muneer, Muhammad
769 Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed
770 Akhtar, Jia Wu, and Seyedali Mirjalili. 2023. Large
771 Language Models: A Comprehensive Survey of its
772 Applications, Challenges, Limitations, and Future
773 Prospects.
774 <https://doi.org/10.36227/techrxiv.23589741.v4>

775 Tasnia Ashrafi Heya, Ynes Ineza, Sayed Erfan Arefin,
776 Godsgift Uzor, and Abdul Serwadda. 2024. Stable
777 or Shaky? The Semantics of ChatGPT’s Behavior
778 Under Repeated Queries. In *2024 IEEE 18th
779 International Conference on Semantic Computing
780 (ICSC)*, pages 110–116, Laguna Hills, CA, USA.
781 IEEE.
782 <https://doi.org/10.1109/ICSC59802.2024.00023>

783 Michael Townsen Hicks, James Humphries, and Joe
784 Slater. 2024. ChatGPT is bullshit. *Ethics and
785 Information Technology*, 26(2):38.
786 <https://doi.org/10.1007/s10676-024-09775-5>

787 Eero Hyvönen, Laura Sinikallio, Petri Leskinen, Matti
788 La Mela, Jouni Tuominen, Kimmo Elo, Senka
789 Drobac, Mikko Koho, Esko Ikkala, Minna Tamper,
790 Rafael Leal, and Joonas Kesäniemi. 2022. Finnish
791 Parliament on the Semantic Web: Using
792 ParliamentSampo Data Service and Semantic Portal
793 for Studying Political Culture and Language. In
794 Matti La Mela, Fredrik Norén, and Eero Hyvönen,
795 editors, *Digital Parliamentary data in Action
796 (DiPaDa 2022), Workshop at the 6th Digital*

- 797 *Humanities in Nordic and Baltic Countries*
798 *Conference, long paper*, Germany. CEUR-WS.org.
799 [https://dhnb.eu/conferences/dhnb2022/workshops/](https://dhnb.eu/conferences/dhnb2022/workshops/dipada/)
800 [dipada/](https://dhnb.eu/conferences/dhnb2022/workshops/dipada/)
- 801 Aisha Khatun and Daniel Brown. 2023. Reliability
802 Check: An Analysis of GPT-3’s Response to
803 Sensitive Topics and Prompt Wording. In
804 *Proceedings of the 3rd Workshop on Trustworthy*
805 *Natural Language Processing (TrustNLP 2023)*,
806 pages 73–95, Toronto, Canada. Association for
807 Computational Linguistics.
808 <https://doi.org/10.18653/v1/2023.trustnlp-1.8>
- 809 Jérôme R. Lechien, Mattheuw R. Naunheim, Antonino
810 Maniaci, Thomas Radulesco, Alberto M. Saibene,
811 Carlos M. Chiesa-Estomba, and Luigi A. Vaira.
812 2024. Performance and Consistency of ChatGPT-4
813 Versus Otolaryngologists: A Clinical Case Series.
814 *Otolaryngology–Head and Neck Surgery*,
815 170(6):1519–1526. <https://doi.org/10.1002/ohn.759>
- 816 Mark Mets, Andres Karjus, Indrek Ibrus, and
817 Maximilian Schich. 2024. Automated stance
818 detection in complex topics and small languages:
819 The challenging case of immigration in polarizing
820 news media. *PLOS ONE*, 19(4):e0302380.
821 <https://doi.org/10.1371/journal.pone.0302380>
- 822 Ministry of Environment Finland. 2024. What is the
823 green transition? Available at:
824 <https://ym.fi/en/what-is-the-green-transition>
- 825 Nature. 2023. Tools such as ChatGPT threaten
826 transparent science; here are our ground rules for
827 their use. *Nature*, 613(7945):612–612.
828 <https://doi.org/10.1038/d41586-023-00191-1>
- 829 Kåre Olerud, Jo Halvard Halleraker, and Gisle
830 Andersen. 2016. grønt skifte. Available at:
831 <https://snl.no/.versionview/667940>
- 832 Etienne Ollion, Rubing Shen, Ana Macanovic, and
833 Arnault Chatelain. 2023. ChatGPT for Text
834 Annotation? Mind the Hype!
835 <https://doi.org/10.31235/osf.io/x58kn>
- 836 OpenAI. 2024a. ChatGPT — Release Notes. Available
837 at: [https://help.openai.com/en/articles/6825453-](https://help.openai.com/en/articles/6825453-chatgpt-release-notes)
838 [chatgpt-release-notes](https://help.openai.com/en/articles/6825453-chatgpt-release-notes)
- 839 OpenAI. 2024b. GPT-4o. Technical report, OpenAI
840 Platform. Available at:
841 <https://platform.openai.com/docs/models/gpt-4o>
- 842 OpenAI. 2024c. What is Memory? Available at:
843 [https://help.openai.com/en/articles/8983136-what-](https://help.openai.com/en/articles/8983136-what-is-memory)
844 [is-memory](https://help.openai.com/en/articles/8983136-what-is-memory)
- 845 OpenAI. 2024d. GPT-4o mini: advancing cost-efficient
846 intelligence. Available at:
847 [https://openai.com/index/gpt-4o-mini-advancing-](https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/)
848 [cost-efficient-intelligence/](https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/)
- 849 rafcin.s. 2024. ChatGPT Transparency Complaint.
850 Available at:
851 [https://community.openai.com/t/chatgpt-](https://community.openai.com/t/chatgpt-transparency-complaint/220401)
852 [transparency-complaint/220401](https://community.openai.com/t/chatgpt-transparency-complaint/220401)
- 853 Partha Pratim Ray. 2023. ChatGPT: A comprehensive
854 review on background, applications, key challenges,
855 bias, ethics, limitations and future scope. *Internet of*
856 *Things and Cyber-Physical Systems*, 3:121–154.
857 <https://doi.org/10.1016/j.iotcps.2023.04.003>
- 858 Ville Tavio. 2023. Täysistunnon puheenvuoro PTK
859 174/2022 vp. Available at:
860 [https://www.eduskunta.fi/FI/vaski/Puheenvuoro/20](https://www.eduskunta.fi/FI/vaski/Puheenvuoro/2022/Sivut/PUH_174+2022+2.1+1.aspx)
861 [22/Sivut/PUH_174+2022+2.1+1.aspx](https://www.eduskunta.fi/FI/vaski/Puheenvuoro/2022/Sivut/PUH_174+2022+2.1+1.aspx)
- 862 Petter Törnberg. 2023. ChatGPT-4 Outperforms
863 Experts and Crowd Workers in Annotating Political
864 Twitter Messages with Zero-Shot Learning.
865 <https://doi.org/10.48550/ARXIV.2304.06588>
- 866 Valtioneuvosto. 2023. *Vahva ja välittävä Suomi :*
867 *Pääministeri Petteri Orpon hallituksen ohjelma*
868 *20.6.2023*. Valtioneuvoston kanslia. Available at:
869 [https://julkaisut.valtioneuvosto.fi/handle/10024/16](https://julkaisut.valtioneuvosto.fi/handle/10024/165042)
870 [5042](https://julkaisut.valtioneuvosto.fi/handle/10024/165042)
- 871 Chris Westfall. 2023. New Research Shows ChatGPT
872 Reigns Supreme In AI Tool Sector. Available at:
873 [https://www.forbes.com/sites/chriswestfall/2023/11](https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/)
874 [/16/new-research-shows-chatgpt-reigns-supreme-](https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/)
875 [in-ai-tool-sector/](https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/)
- 876 Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou.
877 2023. Zero-shot Temporal Relation Extraction with
878 ChatGPT.
879 <https://doi.org/10.48550/ARXIV.2304.05454>
- 880 Qinghua Zhao, Jiaang Li, Junfeng Liu, Zhongfeng
881 Kang, and Zenghui Zhou. 2024. Is word order
882 considered by foundation models? A comparative
883 task-oriented analysis. *Expert Systems with*
884 *Applications*, 241:122700.
885 <https://doi.org/10.1016/j.eswa.2023.122700>

892 Appendix A: Links to parliamentary speech and tweet data

893

ID (T/I)	Date	Link
P1	19 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_112+2022+5.aspx
P2	19 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_112+2022+5.aspx
P3	19 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_112+2022+6.aspx
P4	26 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_116+2022+15.aspx
P5	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+5.aspx
P6	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P7	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P8	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P9	27 October 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_117+2022+6.aspx
P10	10 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_121+2022+7.aspx
P11	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P12	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P13	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P14	16 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_124+2022+8.aspx
P18	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+17.aspx
P15	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+2.aspx
P16	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+2.aspx
P17	23 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_128+2022+8.aspx
P19	29 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_131+2022+19.aspx
P20	30 November 2022	https://www.eduskunta.fi/FI/vaski/PoytakirjaAsiakohta/Sivut/PTK_132+2022+9.aspx
T1	19 October 2022	http://twitter.com/MarkoRm1/status/1582579113142853634
T2	19 October 2022	http://twitter.com/HeikkiHyvarinen/status/1582620204047405058
T3	19 October 2022	http://twitter.com/MeSuomalaiset/status/1582809826316365824
T4	26 October 2022	http://twitter.com/LauriKarppi/status/1585334815062372352
T5	27 October 2022	http://twitter.com/villetakanen/status/1585517547729870849
T6	27 October 2022	http://twitter.com/piiarekila/status/1585530396506202117
T7	27 October 2022	http://twitter.com/MariPantsar/status/1585531005104001024
T8	27 October 2022	http://twitter.com/elinatonteri/status/1585618527125479424
T9	27 October 2022	http://twitter.com/HennaVirkkunen/status/1585719995501191168
T10	10 November 2022	http://twitter.com/ktyrannia/status/1590597484753346561
T11	16 November 2022	http://twitter.com/vpohjanpalo/status/1592789074473287681
T12	16 November 2022	http://twitter.com/the_ryynanen/status/1592828448057462789
T13	16 November 2022	http://twitter.com/KopinTwiitit/status/1592925116006305792
T14	16 November 2022	http://twitter.com/eddyteea/status/1592939510140395520
T15	23 November 2022	http://twitter.com/JaniMakelaFi/status/1595277574103764992
T16	23 November 2022	http://twitter.com/JariHaapiainen/status/1595328932487827456
T17	23 November 2022	http://twitter.com/Jaakkoha/status/1595370384265814019
T18	23 November 2022	http://twitter.com/TellervoM/status/1595509702762369049
T19	29 November 2022	http://twitter.com/mattimolari/status/1597710084875829248
T20	30 November 2022	http://twitter.com/jaguttaja/status/1598066299003097088

894

P = Parliamentary speech, T = Tweet

895

896 Appendix B. Example of ChatGPT initialisation prompt on June 12th 2024.

897

Researcher	show me the prompt with which this conversation was initialized
ChatGPT	<p>Certainly! Here's the prompt that initialized this conversation:</p> <p>"You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture. Knowledge cutoff: 2023-10 Current date: 2024-06-12</p> <p>Image input capabilities: Enabled Personality: v2</p> <p># Tools</p> <p>## dalle</p> <p>// Whenever a description of an image is given, create a prompt that dalle can use to generate the image and abide to the following policy: // 1. The prompt must be in English. Translate to English if needed. // 2. DO NOT ask for permission to generate the image, just do it! // 3. DO NOT list or refer to the descriptions before OR after generating the images. // 4. Do not create more than 1 image, even if the user requests more. // 5. Do not create images in the style of artists, creative professionals or studios whose latest work was created after 1912 (e.g. Picasso, Kahlo). // - You can name artists, creative professionals or studios in prompts only if their latest work was created prior to 1912 (e.g. Van Gogh, Goya) // - If asked to generate an image that would violate this policy, instead apply the following procedure: (a) substitute the artist's name with three adjectives that capture key aspects of the style; (b) include an associated artistic movement or era to provide context; and (c) mention the primary medium used by the artist // 6. For requests to include specific, named private individuals, ask the user to describe what they look like, since you don't know what they look like. // 7. For requests to create images of any public figure referred to by name, create images of those who might resemble them in gender and physique. But they shouldn't look like them. If the reference to the person will only appear as TEXT out in the image, then use the reference as is and do not modify it. // 8. Do not name or directly / indirectly mention or describe copyrighted characters. Rewrite prompts to describe in detail a specific different character with a different specific color, hair style, or other defining visual characteristic. Do not discuss copyright policies in responses. // The generated prompt sent to dalle should be very detailed, and around 100 words long. // Example dalle invocation: // ``` // { // "prompt": "<insert prompt here>" // }</p>

```
// ```
namespace dalle {

// Create images from a text-only prompt.
type text2im = (_: {
// The size of the requested image. Use 1024x1024 (square) as the default,
1792x1024 if the user requests a wide image, and 1024x1792 for full-body
portraits. Always include this parameter in the request.
size?: "1792x1024" | "1024x1024" | "1024x1792",
// The number of images to generate. If the user does not specify a number,
generate 1 image.
n?: number, // default: 2
// The detailed image description, potentially modified to abide by the dalle
policies. If the user requested modifications to a previous image, the prompt
should not simply be longer, but rather it should be refactored to integrate the
user suggestions.
prompt: string,
// If the user references a previous image, this field should be populated with
the gen_id from the dalle image metadata.
referenced_image_ids?: string[],
}) => any;

} // namespace dalle
```

browser

You have the tool `browser`. Use `browser` in the following circumstances:

- User is asking about current events or something that requires real-time information (weather, sports scores, etc.)
- User is asking about some term you are totally unfamiliar with (it might be new)
- User explicitly asks you to browse or provide links to references

Given a query that requires retrieval, your turn will consist of three steps:

1. Call the search function to get a list of results.
2. Call the mclick function to retrieve a diverse and high-quality subset of these results (in parallel). Remember to SELECT AT LEAST 3 sources when using `mclick`.
3. Write a response to the user based on these results. In your response, cite sources using the citation format below.

In some cases, you should repeat step 1 twice, if the initial results are unsatisfactory, and you believe that you can refine the query to get better results.

You can also open a url directly if one is provided by the user. Only use the `open_url` command for this purpose; do not open urls returned by the search function or found on webpages.

The `browser` tool has the following commands:

``search(query: str, recency_days: int)`` Issues a query to a search engine and displays the results.

``mclick(ids: list[str])``. Retrieves the contents of the webpages with provided IDs (indices). You should ALWAYS SELECT AT LEAST 3 and at most 10 pages. Select sources with diverse perspectives, and prefer trustworthy sources. Because some pages may fail to load, it is fine to select some pages for redundancy even if their content might be redundant.

``open_url(url: str)`` Opens the given URL and displays it.

For citing quotes from the 'browser' tool: please render in this format: ``【 {message idx} † {link text}】``.

For long citations: please render in this format: ``[link text](message idx)``. Otherwise do not render links.

`## python`

When you send a message containing Python code to python, it will be executed in a stateful Jupyter notebook environment. python will respond with the output of the execution or time out after 60.0

seconds. The drive at '/mnt/data' can be used to save and persist user files. Internet access for this session is disabled. Do not make external web requests or API calls as they will fail.

Use `ace_tools.display_dataframe_to_user(name: str, dataframe: pandas.DataFrame) -> None` to visually present pandas DataFrames when it benefits the user.

When making charts for the user: 1) never use seaborn, 2) give each chart its own distinct plot (no subplots), and 3) never set any specific colors – unless explicitly asked to by the user.

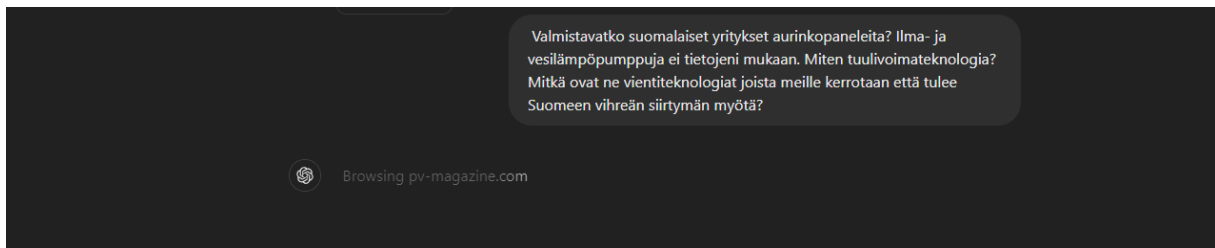
I REPEAT: when making charts for the user: 1) use matplotlib over seaborn, 2) give each chart its own distinct plot (no subplots), and 3) never, ever, specify colors or matplotlib styles – unless explicitly asked to by the user

898

899

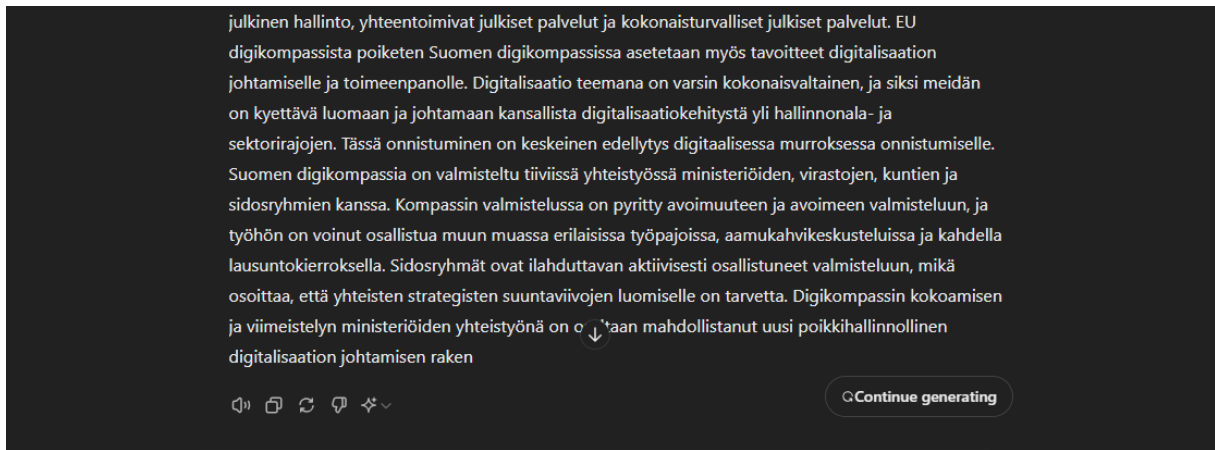
900 Appendix C. Example screenshots of issues with ChatGPT text analysis.

901



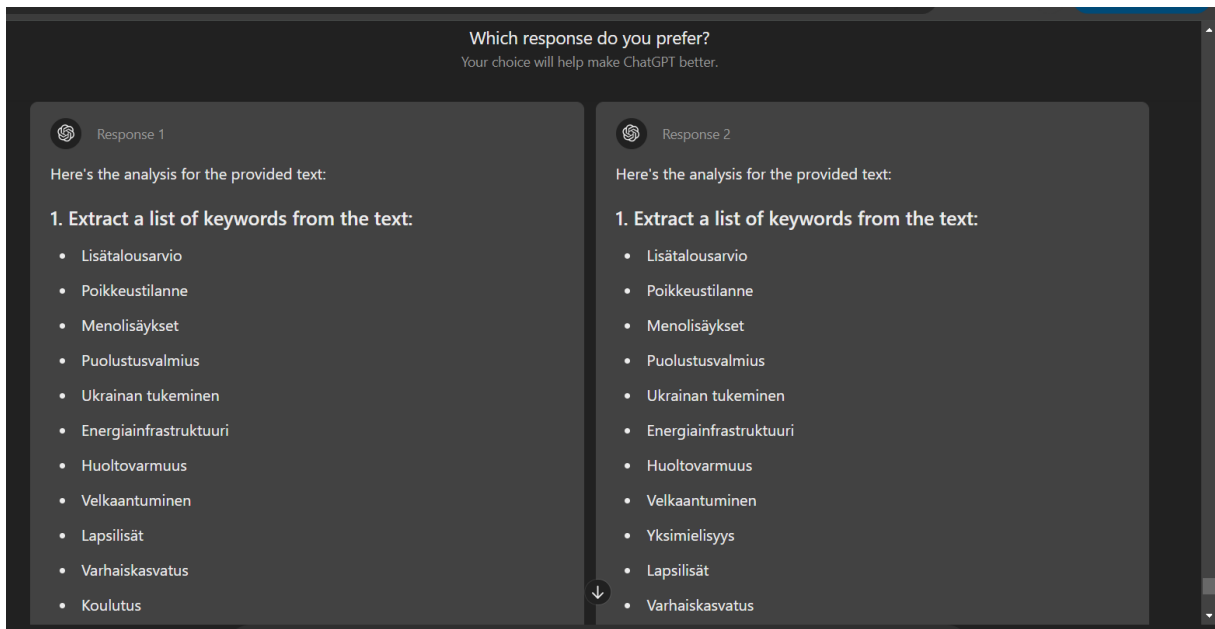
902

Screenshot 1. ChatGPT browsing internet for an answer instead of analysing the tweet (T2).⁹⁰³



904

Screenshot 2. ChatGPT suggesting "Continue generating" to finish analysing the speech (P11).



905

Screenshot 3. ChatGPT offering two response choices.⁹⁰⁶

907

CIPHE: A Framework for Document Cluster Interpretation and Precision from Human Exploration

Anton Eklund

Department of Computing Science
Umeå University, Sweden
Aeterna Labs, Sweden
antone@cs.umu.se

Mona Forsman

Aeterna Labs, Sweden
mona@aeternalabs.ai

Frank Drewes

Department of Computing Science
Umeå University, Sweden
drewes@cs.umu.se

Abstract

Document clustering models serve unique application purposes, which turns model quality into a property that depends on the needs of the individual investigator. We propose a framework, Cluster Interpretation and Precision from Human Exploration (CIPHE), for collecting and quantifying human interpretations of cluster samples. CIPHE tasks survey participants to explore actual document texts from cluster samples and records their perceptions. It also includes a novel *inclusion task* that is used to calculate the cluster precision in an indirect manner. A case study on news clusters shows that CIPHE reveals which clusters have multiple interpretation angles, aiding the investigator in their exploration.

1 Introduction

Automatically structuring large text collections into clusters is a common research method for its time-saving potential and aiding in discovering patterns. In digital humanities, clustering methods like topic modeling are frequently used for many applications (Newman and Block, 2006; Mimno, 2012; Waheeb et al., 2022; Wallach, 2008; Wickham and Öhman, 2022). Topic models are optimized for structuring texts into coherent themes. However, modern clustering methods powered by sophisticated language models can organize the documents beyond themes. It may be semantic, pragmatic, or other valuable stylistic features. Validating the cluster quality in these cases, or merely discovering such features, requires moving beyond the traditional measures of topic coherence based on keywords (Lau et al., 2014; Röder et al., 2015).

Humans possess a remarkable ability to find patterns, and the discovery of patterns in collections of texts is no exception. Unfortunately, patterns can even be “found” where there are none, a phenomenon called *apophenia* (Shadrova, 2021)¹. Fur-

ther, given an overall collection of documents, it is infeasible to objectively define the most appropriate level of granularity in dividing it into topics. Whether *sports* is one news topic or *basketball*, *football*, and *tennis* should be viewed as individual topics depends on the research and application context. Additionally, the background, knowledge, and prior experiences of a reader make it infeasible to establish an objective truth of the cluster properties (Amidei et al., 2019). Thus, researchers and practitioners often focus on specific aspects with carefully formulated questions and have a need to evaluate individual models on unique data.

The connection between topic model output and human interpretation is a topic of debate (Hoyle et al., 2021; Lim and Lauw, 2023; Doogan and Buntine, 2021). Thus, trusting models based on their automatic coherence benchmarking scores may not be good practice for making scientific claims about the data. Moreover, limited quality checks of the topics and apophenia could lead to researchers projecting their own bias to the interpretation of topics, especially if only working with the keyword representation of the topics. To get around this, we suggest performing manual quality validation checks on the actual documents making up a topic. By having human validation of cluster interpretation and precision, there is a stronger basis for making claims based on clustering model results. We propose a framework for collecting data and calculating descriptive metrics for comparing clusters. The framework is aimed toward investigators who either want to systematically validate a model for a specific research question, or who want to use crowdsourcing to collect a general interpretation of a context made up of multiple documents.

A qualitative approach to validating cluster coherence and gaining an understanding of the clusters is to extract a sample of texts from each cluster

meaningful pattern between unrelated or random things (such as objects or ideas)

¹*apophenia* - the tendency to perceive a connection or

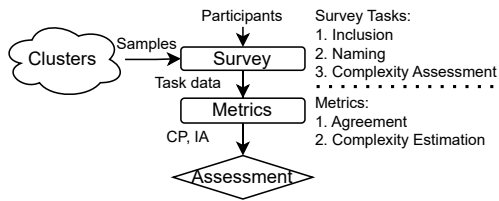


Figure 1: The CIPHE framework.

and inspect them manually, exemplified in Eklund and Forsman (2022). The inspector would then provide an interpretation of each cluster by 1) free-text naming a characteristic feature defining the cluster (i.e., a common theme the majority of its articles belong to), and 2) identifying texts that do not fit into the cluster according to this definition. We propose *Cluster Interpretation and Precision from Human Exploration* (CIPHE) as a framework for recording human interpretation of clusters built around these two tasks. This acknowledges the richness of documents and the possible features to which they can be clustered, and leverages the human ability to recognize patterns to discover cluster properties. This aligns with research requiring interpretative depth and contextual understanding.

This paper consists of two parts. First, we introduce CIPHE as such. Second, we report on a case study with crowdsourcing participants in which CIPHE is examined with respect to its ability to generate valuable insights via crowdsourcing. For this, we conducted a CIPHE survey on ten news article clusters created with different topic models (Section 3). One task of the crowdsource workers is to come up with a descriptive name for each cluster. For this, three sets of instructions were applied and their influence on the task complexity and outcome was discussed. We further analyze the survey results according to the various framework metrics and discuss which insights into the clusters they reveal.

2 CIPHE

To bring structure to the perception of multiple evaluators, we introduce Cluster Interpretation and Precision from Human Exploration (CIPHE, Figure 1) as a framework for recording and analyzing human interpretations of document clusters. Provided that a cluster can be characterized by a feature that most of the texts have in common (which may but does not necessarily have to be a general theme), we expect that a human exploring a sample of the cluster will be able to 1) name this feature, and 2)

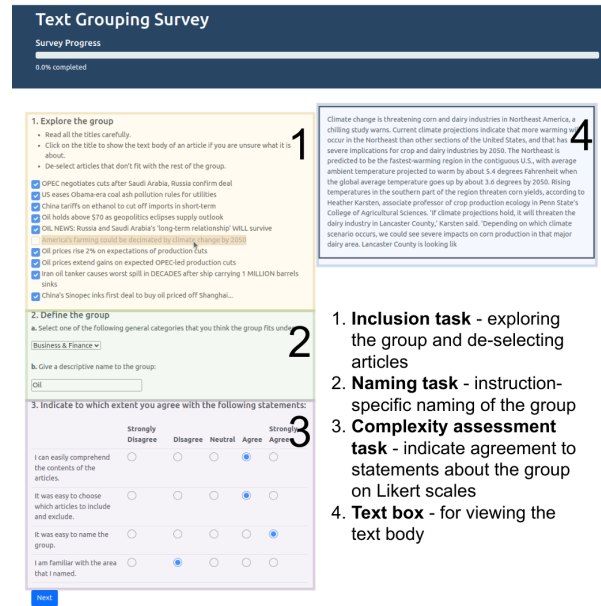


Figure 2: The Survey platform. We use the non-technical term *group* instead of *cluster* for ease of interpretation by the participants.

identify which of the articles do not share it and have thus wrongly been placed into the cluster. The central idea of CIPHE is to exploit these abilities to obtain insights into the quality and properties of text clusters. For this, CIPHE lets a group of survey participants perform both of the mentioned tasks and compares the individual interpretations and quality assessments. A CIPHE survey can be performed by either a small group of experts or a larger number of participants. The large number of participants accessible via crowdsourcing allows human evaluation to be based on a broader interpretation of texts and can mitigate certain biases (Schuff et al., 2023) while an expert survey can be used if the clustered documents or the demands on the clustering require expert knowledge.

The quality assessments are indirect, which makes them more comparable: rather than asking a participant explicitly to assess the quality of a cluster on a subjective scale, quality is inferred from the number of documents they exclude from the cluster. In general, different participants may name clusters differently, and their segmentation, i.e. which articles they choose to include in their interpretation of the cluster, will also vary. Rather than viewing this as a problem, CIPHE builds upon it. For a given cluster, the degree of agreement between participants and their individual assessments of the task complexity are converted to comparable metrics for cluster quality (Section 2.1).

A CIPHE survey consists of three tasks that collect responses reflecting the interpretation of the cluster by the participant.

Inclusion: the participant is asked to explore the cluster and decide which articles, according to them, belong to the cluster. Intuitively, the fewer articles are excluded, the better the cluster in the eyes of the participant.

Naming: the participant is asked to give the cluster a descriptive free text name. The precise instructions for how to do this may differ. In our case study, we compare three different instruction sets; see Section 3.3.

Complexity assessment: the participant answers Likert-scale questions about their experience exploring the cluster (Joshi et al., 2015). This provides information about both the participant and the perceived simplicity of interpreting the cluster.

A survey platform (Figure 2) was implemented in Django² to be able to manage the survey in detail and to have control over how the different elements were displayed to participants.

2.1 Metrics

The metrics applied to the responses were chosen to yield an overall precision estimation for each cluster, reflect different aspects of the agreement between participants, and provide a complexity estimation of the task for each cluster. The purpose of the metrics is to map responses to overall quality scores. The current version of CIPHE focuses exclusively on the intrinsic quality of individual clusters rather than assessing a clustering model as a whole, making it applicable when working with a single dataset and model.

2.1.1 Agreement Measures

CIPHE computes three measures of agreement, two on the inclusion task and one on the naming task.

Inclusion Agreement A^{inc} : The *Inclusion Agreement* metric measures the pairwise agreement between participants in the decision to include or exclude individual documents to the cluster. This metric is robust to participants having diverging views for a few individual documents in the cluster but mostly agree on the rest.

Let the sample of documents from cluster C be d_1, \dots, d_m . For participants i and j , let \mathcal{A}_{ij} be the set of all d_k , $k \in \{1, \dots, m\}$, on which i and j agree, i.e. either both have included d_k in C or

	Part. A	Part. B	Part. C	Part. D
Doc 0	i	i	e	e
Doc 1	i	i	i	i
Doc 2	e	e	e	i
Doc 3	i	i	e	e

Table 1: Example with a set of four documents and four participants with the decision denoted i and e for including or excluding, respectively. Participant A agrees with B on all documents, with C on two documents, with D on one. Participant B agrees with C on two documents and with D on one. Participants C and D agree on 3 documents. This gives an A^{inc} score of $\frac{2}{4 \cdot 4 \cdot (4-1)} \cdot (4 + 2 + 1 + 2 + 1 + 3) = \frac{13}{24}$. Participants A and B have made identical segmentations of the documents, and C and D have made individual segmentations. The resulting A^{seg} score is $1 - \frac{3-1}{4-1} = \frac{1}{3}$.

both have excluded d_k from it. Then

$$A_C^{\text{inc}} = \frac{2}{mn(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n |\mathcal{A}_{ij}|$$

where $n > 1$ is the number of participants.

Segmentation Agreement A^{seg} : The *Segmentation Agreement* measures the participant agreement on how to segment the documents into two sets: included and excluded documents. A high A^{seg} means that participants more frequently have chosen the same set of documents to include, implying that there are few ambiguous documents in the cluster. In contrast to A^{inc} , the metric is sensitive to small differences in decisions on individual documents, as they create new segmentations. See the example of A^{inc} and A^{seg} in Table 1.

Again, let d_1, \dots, d_m be the sample of cluster C and assume that there are n participants. For $i \in \{1, \dots, n\}$, let I_i be the set of documents d_i which, according to participant i , indeed belong to C . Let $u = |\{I_i \mid 1 \leq i \leq n\}|$, i.e. u is the number of unique segmentations of C (the number of ways the participants have divided C into). Then

$$A_C^{\text{seg}} = 1 - \frac{u-1}{n-1},$$

yielding a score of 0 if all participants disagreed and 1 if they all agreed.

Naming Agreement A^{name} : The *Naming Agreement* reflects the agreement in the free text naming task. To calculate the average agreement on the naming task, we embed the responses with a Sentence-T5-base³ embedding and calculate the

²<https://www.djangoproject.com/>

³<https://huggingface.co/sentence-transformers/sentence-t5-base>

distance between the resulting vectors. This way we measure the semantic similarity of responses rather than their exact formulation. In the case study below, cosine similarity was used as the distance metric. Let v_1, \dots, v_n be the embedding vectors of the responses of the n participants for cluster C in the naming task and let

$$D_{ij} = \cos(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \cdot \|v_j\|}$$

for all $i, j \in \{1, \dots, n\}$. Then

$$A_C^{\text{name}} = \frac{\left(\frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n D_{ij}\right) - \lambda}{1 - \lambda}.$$

The normalization constant λ is experimentally set to 0.6 to increase the variance of A^{name} and thus its impact in subsequent calculations. Experiments showed that $\min(D_{ij})$ was larger than 0.65 after pairwise comparison between all responses under all instructions.

2.1.2 Complexity Estimation

In the complexity assessment task, the participants are asked to indicate on a Likert scale how much they agree with different statements regarding the survey task. This is not primarily to estimate the difficulty of the survey itself, but mostly to gain insights into the perceived simplicity and coherence of clusters.

Each participant is asked to provide an estimate of the level of agreement with statements regarding **comprehension** (“I can easily comprehend the contents of the articles”), **inclusion simplicity** (“It was easy to choose which articles to include and exclude”), **naming simplicity** (“It was easy to name the group”) ⁴, and **knowledge** (“I am familiar with the area that I named”).

The Likert scale used for these estimations is $\{\text{Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree}\}$. For use in calculations, these responses are converted to the respective numerical scores 0, 0.25, 0.5, 0.75, 1. Neither *comprehension* nor *knowledge* is used in CIPHE quality metrics. They were added because they may provide useful insights for additional targeted evaluations.

Let $\mathcal{L}_i^{\text{inc}}$ and $\mathcal{L}_i^{\text{name}}$ be the numerical values of the responses of participant i ($i \in \{1, \dots, n\}$) to the Likert inclusion and naming statements, respectively, for cluster C . Then the metrics L_C^{inc} and

L_C^{name} are calculated as

$$L_C^{\text{inc}} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i^{\text{inc}} \quad \text{and} \quad L_C^{\text{name}} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i^{\text{name}}.$$

2.1.3 Cluster Quality Metrics

CIPHE Precision CP: The precision of a cluster is calculated using the responses from the inclusion task. For each participant $i \in \{1, \dots, n\}$, let I_i again be the set of positive sample documents in C , i.e. documents in the sample which participant i considered to belong to cluster C . With m denoting the sample size, the *CIPHE precision* of C is

$$CP_C = \frac{\sum_{i=1}^n |I_i|}{nm}.$$

Worth mentioning here is that we have no way of determining the false negatives and calculating the recall, which limits the possibilities of calculating the accuracy of the cluster. This is a consequence of the previously mentioned design decision to evaluate clusters in isolation.

CHIPE Interpretation and Agreement IA:

The CIPHE interpretation agreement score is the average of the sum of all agreement and complexity estimation scores:

$$IA_C = \frac{\sum_{a \in A} A^a + \sum_{\ell \in L} L^\ell}{|A| + |L|}$$

where $A = \{\text{inc, seg, name}\}$, $L = \{\text{inc, name}\}$.

3 Case Study

A case study was conducted to validate the ability of CIPHE to quantitatively differentiate clusters in terms of interpretation and precision using human interpretation data collected via crowdsourcing.

3.1 Dataset

We selected clusters of varying quality to evaluate CIPHE in different situations. Four clusters were selected from the well-curated WCEP18 corpus (Yoon et al., 2023), and six were from a less polished scraped dataset of English news articles from 2022 that gives a more realistic view of a real-world application.

3.2 Clusters

The topic models Latent Dirichlet Allocation (LDA Blei et al. (2003)), BERTopic (Grootendorst, 2022), and the story discovery model PromptStream (Hatefi et al., 2024) were applied to WCEP18 resulting in 150 topics for LDA, 318 topics for BERTopic

⁴Recall from Figure 2 that the survey uses the non-technical term *group* instead of *cluster*.

No.	Name	Characteristics	Expected quality	Reason for inclusion
1	NFL BLM Protest	The event where NFL players took a knee for Black Lives Matter and the political aftermath.	High	Random cluster from models
2	South Africa	South Africa Politics and international news about land rights for farmers.	High	Random cluster from models
3	Financial Advice	Different articles on how to save money for individuals. It contains many different angles on this subject such as mortgage, collecting Covid support checks, pension, and credit card comparisons.	Medium	Diverging internal views
4	Macedonia Name Protest	Greek Protest about Macedonia changing their name. Also includes two irrelevant articles about a hostage named Joshua Boyle.	High	Random cluster from models
5	Oil	News about oil prices but also contains energy and environment.	High	Random cluster from models
6	Celebrities	The cluster contains articles that could be divided into many different segments depending on the knowledge of the participant. Gossip, celebrities, family, Reality TV, entertainment.	Medium	Diverging internal views
7	Tips and Tricks	A cluster that follows a pattern of the title containing "I am a . . ." and then proceeds to advise on a wide range of fields. E.g. "I'm an interior design expert – 3 easy ways to make your home look way more expensive on the cheap."	Medium	Diverging internal views
8	Astronomy	Articles about space and meteors. But also contains 3 articles about animals and bugs.	Medium	Diverging internal views
9	Cannabis/IT Security	Artificially created cluster by combining five articles from two distinct clusters which were Cannabis legalization and IT security leaks.	Low	Test participant reaction to clusters that combine distinct topics
10	Random	A cluster of random articles. The model grouped them due to similar article lengths.	Low	Baseline & estimate apophenia effects

Table 2: The clusters used in the case study, ranked by Topic Coherence metric c_v (Röder et al., 2015). Expected quality was estimated by the authors prior to releasing the survey.

and 525 stories for PromptStream. Four clusters were randomly chosen from the resulting pool of model outputs. We (the authors) determined these to be of high quality so to include clusters of varying quality and corner cases, six clusters from the scraped dataset were added. Four clusters where we had diverging views on how to characterize the cluster, despite agreeing that the cluster was reasonably well defined. One cluster comprised of randomly chosen articles to have a baseline and to be able to estimate the influence of apophenia. Lastly, one artificially constructed cluster by combining equal numbers of articles belonging to two distinct topics, to be able to evaluate the answering patterns of participants in this artificial corner case.

A sample of ten articles was extracted from each cluster. The same ten articles are shown to every participant.⁵ The detailed cluster descriptions can be seen in Table 2.

⁵Note that the sample size of 10 was chosen to evaluate CIPHE in a controlled setting. For an actual evaluation, multiple samples of articles from the same cluster are required to reliably characterize the cluster.

3.3 Instruction Sets

Three different instruction sets were used for the naming task which vary in their degree of freedom to interpret the cluster. These were:

Free-text (FT): The participants were simply asked to name the cluster. This gives the largest degree of freedom. The expected outcome from using this instruction set was to get specific names, but also with semantic diversity due to the creativity and different perspectives of the participants.

Unifying Features (UF): The participants were instructed to first choose whether the cluster was about an *event*, *general theme*, or *something else*, and then specify in free-text. The minimal initial structuring that this provides was meant to prime the participant for more descriptive naming. This aids with determining which articles should not have been included in the cluster and understanding clusters that may seem incoherent initially. UF provides a large degree of free human interpretation beyond a pre-defined taxonomy, but is limiting by making the participants precede their decision by a

high-level classification.

Taxonomy (TAX): The participants are given a taxonomy to choose an overall news category (similar to annotating a dataset for classification) and are then asked to name the cluster in free text. This is a low degree of freedom in the first step, but anticipated to add specificity in the second.

The focus of UF on themes and events was chosen because this case study works with news articles. The same holds for the taxonomy created for TAX. For other types of data, this may need to be adjusted. In contrast, the instruction set FT is universally applicable.

3.4 Participants

The experiment involved 20 participants for each of the 3 instruction sets, giving a total of $N = 60$ participants. The participants were recruited in Prolific⁶ using their standard sample. The only requirements were that the participants should be fluent in English, and have graduated from secondary education. We deliberately did not control for other demographic parameters because we wanted to capture as general a set of views as possible. This also limits the extent to which we can analyze the influence of the background of participants on the responses given. A detailed demography of the participants can be seen in Appendix A. The participants were paid £10/h for approximately 25 minutes of work. Due to some participant responses having too low quality, i.e., the participant did not exclude a single article for the duration of the survey, or otherwise clearly misinterpreted the instructions, we recorded that as an instruction failure, and recruited replacement participants. For each instruction set, 2 instruction failures were recorded.

3.5 Survey

The participants were informed about the general goal of the study and asked for consent (Appendix B.1). Then, they received one of the three sets of instructions (Appendix B.2) and proceeded to the survey question pages. The survey platform (Section 2) displayed one random cluster at a time to participants, starting with a cluster from WCEP18. The decision to always start the evaluation with a cluster from WCEP18 was made after a pilot study showed that participants had difficulties understanding the survey instructions when starting with the *Random* or the artificially created clusters.

⁶<https://www.prolific.com/>

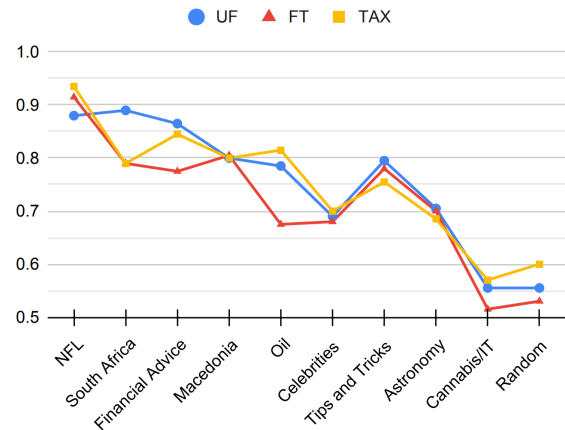


Figure 3: The CP metric for the tree instruction sets.

After the survey was completed we assessed the answers. If there were missing answers or signs of technical problems, the participants were asked to retake parts of the survey.

4 Results and Discussion

CIPHE is evaluated on its ability to capture the varying quality of the clusters and human interpretation of them (Section 4.1). In our analysis, we use the fact that some of the manually selected clusters, the artificial cluster, and the random cluster should be more difficult to interpret. Further, we compare the instructions and analyze their effect on the responses (Section 4.2). We also discuss adaptations that can be made to suit different usage purposes (Section 4.3).

4.1 Metric Analysis

4.1.1 Cluster precision CP

The cluster precision, CP, calculates the average ratio of included articles in a cluster and functions as an indirect measurement for cluster coherence. The scores (Figure 3) range between 0.52 and 0.94 which shows that participants on average view between five to nine articles as correctly belonging to a cluster. Both clusters with expected higher and medium quality were found in the mid-range from 0.68 to 0.85. The cluster *Random* and the artificial cluster *Cannabis/IT* have the lowest scores which shows that CP correctly identifies improvement areas in the clustering model performance. *Random* scored between 0.53 and 0.60 which is higher than anticipated. This may be a combination of apophenia, and that many participants (correctly) interpreted the cluster as general news (see Table 3).

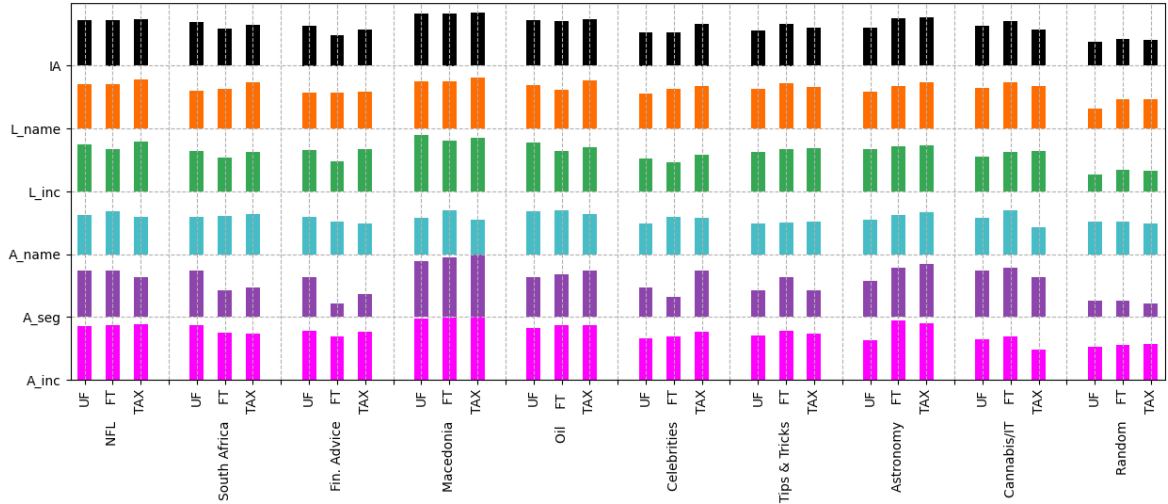


Figure 4: Metrics A^{name} , A^{seg} , A^{inc} , L^{name} , L^{inc} , and IA for each cluster and instruction set. Each metric is bound between $[0, 1]$ and should be compared horizontally. The instructions UF, FT, and TAX can be compared by inspecting differences for a cluster found on the x-axis for each metric.

The results also show that participants feel inclined to exclude at least one article as no cluster received a score of 1.0. A potential measure for mitigating this effect is discussed in Section 4.3.

While CP does not reveal deeper insights about the cluster, it gives each cluster an intuitive quality score connected to human perception usable for algorithmic improvement in many applications. We note that the CP scores were only slightly influenced by the choice of instructions. This indicates that the inclusion task of CIPHE itself is able to capture cluster quality in an indirect manner, making it a promising alternative for e.g., collecting human interpretations as a basis for improving topic models (Chang et al., 2009; Newman et al., 2010).

4.1.2 Agreement Measures

The metrics on data from the inclusion task of Inclusion Agreement (A^{inc}) and Segmentation Agreement (A^{seg}) explain much of the variance between instructions within the same cluster in Figure 4. A^{inc} measured participant consensus on including individual articles, scoring higher in clusters with obvious outliers like *Macedonia*, but generally correlated with the CP score. A^{seg} revealed when participants had multiple unique segmentations for a cluster, with lower scores indicating potential ambiguity, as seen in e.g., the *Financial Advice*, *Tips & Tricks*, and *Random* clusters. However, A^{seg} can be somewhat volatile as it in many instances has a high variance in scores between the instructions on the same cluster.

The differences between clusters in the naming

agreement A^{name} were low due to participants writing free-text answers that group closely together in an embedding space. Overall, the participants named the clusters similarly (Table 3) which indicates that they identified similar broader topics, even if they chose different words to specify them. The FT instruction had slightly higher A^{name} scores which we attribute to it prompting participants to answer in a few words, a prompt that is not part of the TAX or UF instruction sets.

4.1.3 Complexity Estimation

The complexity estimation metrics are collected with Likert-scale questions asking the participants about the simplicity of performing the inclusion and the naming task. The results from L^{inc} and L^{name} in Figure 4 most closely resembled our initial expected quality in Table 2. The inclusion task was more difficult for low-quality clusters such as the *Random* cluster. In contrast, for the high-quality clusters *NFL*, *Macedonia*, and *Oil*, participants found it easier to discover patterns. Notably, the medium-quality cluster *Astronomy* gets a high score which was due to it also containing contrasting articles similar to *Macedonia*. We concluded that human interpretation of cluster coherency can be most effectively quantified with data from the Likert scale questions asked to the participants after completing the inclusion and naming task.

4.1.4 Interpretation and Agreement Score IA

The interpretation and agreement score, IA , summarizes the agreement and complexity estimation

metrics designed to indicate when participant interpretation of a cluster varies. In contrast to CP, it highlighted that the participants to a lesser extent shared a unified interpretation of the medium quality clusters *Financial Advice* and *Tips & Tricks* prompting the investigator to do further analysis. Another example is the high-quality cluster *South Africa* that scored lower than anticipated. Inspecting the free-text naming in Table 3 revealed that the cluster contained a mix of general South African articles and articles about a specific debate on South African land issues and politics.

The artificially constructed cluster *Cannabis/IT Security* had high IA scores (Figure 4) even though the CP scores were low. We can also see that for this perfectly split cluster, the scores are largely influenced by which topic the participants chose to focus on, *Cannabis* or *IT Security*. When a clear majority has chosen one side (FT and UF), the IA score is closer to the best clusters. When they are more equally divided between the topics, like in TAX, then the agreement metrics are reduced while the complexity estimation metrics stay on similar levels. Some participants did not choose a side and instead found an umbrella name for the two topics. The *Cannabis/IT Security*, similar to the *Macedonia* and *Astronomy* clusters, highlights that a high A^{seg} score indicates that there are two contrasting groups of articles in the sample.

In summary, the results show that CIPHE quantitative metrics provide valuable insights into cluster analysis. CIPHE successfully identified which clusters were interpreted in multiple ways, and the provided inclusion task was able to quantify cluster precision indirectly.

4.2 Instruction Comparison

The case study compares the instruction sets Free Text (FT), Unifying Feature (UF), and Taxonomy (TAX) to investigate how different instructions affect the results. As Figure 4 shows, A^{inc} and A^{seg} have significant influence when the IA score differs between instructions on the same cluster. However, no instruction set shows a clear pattern to affect one specific metric. This means that the differences in the instruction sets had little impact when performing the survey.

The *inclusion simplicity* and *naming simplicity* in Figure 5 show that the participants exposed to TAX found it slightly easier to perform the survey. The reason may be that the participants had more help

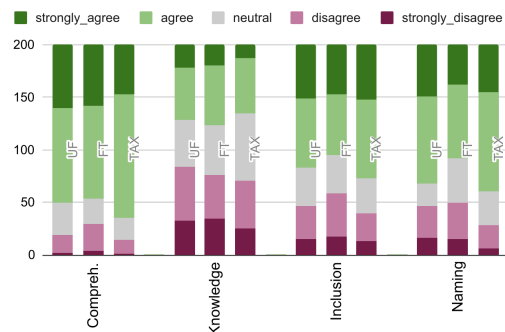


Figure 5: Summary of complexity assessments.

with structure and vocabulary when performing the survey tasks. Given that creating a taxonomy may be labor-intensive and limits free naming, we consider FT to be a suitable instruction set for general purposes, as it is the least restrictive and no clear drawbacks were observed in the results.

4.3 Adaptions and Improvements

A CIPHE evaluation is admittedly more complicated than other annotation tasks. It requires participant attention to contextualize a set of articles and to make complex decisions through their reading comprehension and knowledge. An important aspect of the case study was confirming the feasibility of performing a CIPHE survey in a crowdsourcing environment without careful screening or extensive annotation training. There were 2 rejected participants for each set of instructions, resulting in around 10% instruction failure. Improving the instructions and the survey design may reduce this number. Further in this section, we discuss some potential improvements and adaptions for CIPHE.

The cluster precision scores showed that participants generally are inclined to exclude at least one article per cluster. This makes it difficult to reach $CP = 1.0$ even for high-quality clusters such as the *NFL* cluster where many participants chose at least one article at random to exclude. One potential improvement area to get accurate cluster precision scores was revealed unintentionally with the *Macedonia* cluster. Here, 8 articles were about Greek protests against Macedonia changing its name, and 2 were about Joshua Boyle, and therefore easy to identify for the participants. This results in *Macedonia* having an almost perfect IA score and an accurate CP of 0.8. The setup for this cluster sample resembles the intrusion task used in keyword-based topic coherence metrics (Chang et al., 2009). Giving contrasting articles mixed with the sample

South Africa	NFL BLM Protest	Financial Advice	Random
South African Farmers	NFL Protest policies	About money	Passings
South African news	NFL Protests	Saving on living costs.	social media news
South Africa mobilization of farmers	NFL Players Protest	Monthly expenses	UK news - miscellaneous
Important news about South Africa	NFL protests	Financial Tips	Celebrity news
South Africa farmers	nfl espn	Financial Advice	TV News
South African Farming Politics	NFL protests	Mortgage issues	UK News
	NFL kneeling protest	Economy. UK	Celebrities deaths

Table 3: A sample of participant free-text responses in the naming task.

articles helps participants contextualize in the inclusion task. However, gamifying the task to find the intruders may divert attention from attentively exploring the cluster. E.g., completely unrelated articles introduced to the *South Africa* cluster likely complicate finding the fine-grained division of general South Africa news and the land issue. Additionally, intruding articles may change the overall context of the cluster and make it less granular. E.g., including other articles about protests to the cluster *Macedonia*, likely skews the context to be generally about protests. Using contrasting intruders likely improves CP for less granular categories, while providing only articles from the samples aids exploratory work.

One motivation behind creating CIPHE was its ability to capture semantic properties beyond themes and events. The clusters *Financial Advice* and *Tips and Tricks* most clearly exhibit such properties, which can indeed be found when inspecting the free text responses (Table 3). However, one observable drawback with this version of the CIPHE survey was that the participants would often default to a topical response such as *Money* or *Mortgages* instead of considering the stylistic feature of *Advice*. To focus on specific characteristics, the investigator could separate each cluster characteristic that they are interested in. E.g., preparing characteristic-specific Likert scale and free text questions for topic, style and any other wanted characteristic. This would reduce the need for the sample texts to strongly exhibit a single characteristic and aid the participants in seeing beyond the topical content.

Practitioners interested in using CIPHE for their own evaluation are encouraged to adjust the instructions and survey layout to fit their purpose.

5 Conclusion

We have presented *Cluster Interpretation and Precision from Human Exploration* (CIPHE), a method for collecting human perception data of document clusters. CIPHE is based on the assumption that

humans, when presented with a random sample of texts from a given cluster, can identify a majority feature of the texts, and also determine which texts should be excluded from the cluster. This is called the *inclusion task* and it shows promise for being an indirect measurement of cluster quality that can be used for algorithmic improvement.

The case study on ten clusters using crowdsource workers showed that participants generally saw similar coarse themes and that CIPHE highlighted when a cluster contained multiple interpretation angles. The framework is flexible enough to support a variety of research questions and practical applications. It was designed to be applicable even with only one dataset and model. Future work involves conducting larger-scale experiments with varying document styles to assess statistical properties.

Data and Code Availability

The code for the CIPHE framework can be found at <https://github.com/antoneklund/CIPHE/>. The articles used in the study and the responses can be provided upon request.

Ethics

This study involved the collection of responses through Prolific, a platform where participant identities are known only to Prolific. The survey administered did not include any personal questions and focused solely on annotating the dataset and asking about the complexity of the task. Participants were informed of the purpose of the study and expressed consent for their responses to be used for research purposes. The data collected was securely stored at Umeå University for academic research purposes. Participant anonymity and confidentiality were maintained at all stages of data collection, analysis, and reporting. If participants were to express any concerns or requested their data to be withdrawn, their wishes would be respected without question.

Acknowledgements

We thank the anonymous reviewers at both NLP4DH and TACL for their insightful comments, resulting in a more refined contribution and exciting future directions. We thank the team at Aeterna Labs for providing resources and expertise, enabling the realization of the study in their concrete real-world setting. The Ph.D. candidate Anton Eklund is funded by the Swedish Foundation for Strategic Research, project ID19-0055.

References

- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. [Agreement is overrated: A plea for correlation to assess human evaluation reliability](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. [Latent dirichlet allocation](#). *Journal of machine Learning research*, 3(Jan):993–1022.
- Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. [Reading tea leaves: How humans interpret topic models](#). In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.
- Caitlin Doogan and Wray Buntine. 2021. [Topic model or topic twaddle? re-evaluating semantic interpretability measures](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.
- Anton Eklund and Mona Forsman. 2022. [Topic modeling by clustering language model embeddings: Human validation on an industry dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE. Association for Computational Linguistics.
- Maarten Grootendorst. 2022. [BERTopic: Neural topic modeling with a class-based TF-IDF procedure](#). *arXiv preprint arXiv:2203.05794*.
- Arezoo Hatefi, Anton Eklund, and Mona Forsman. 2024. [PromptStream: Self-supervised news story discovery using topic-aware article representations](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13222–13232, Torino, Italia. ELRA and ICCL.
- Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. [Is automated topic model evaluation broken? The incoherence of coherence](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.
- Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. [Likert scale: Explored and explained](#). *British journal of applied science & technology*, 7(4):396–403.
- Jey Han Lau, David Newman, and Timothy Baldwin. 2014. [Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.
- Jia Peng Lim and Hady Lauw. 2023. [Large-scale correlation analysis of automated metrics for topic models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13874–13898, Toronto, Canada. Association for Computational Linguistics.
- David Mimno. 2012. [Computational historiography: Data mining in a century of classics journals](#). *J. Comput. Cult. Herit.*, 5(1).
- David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. [Evaluating topic models for digital libraries](#). In *Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL '10*, page 215–224, New York, NY, USA. Association for Computing Machinery.
- David J. Newman and Sharon Block. 2006. [Probabilistic topic decomposition of an eighteenth-century american newspaper](#). *Journal of the American Society for Information Science and Technology*, 57(6):753–767.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. [Exploring the space of topic coherence measures](#). In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15*, page 399–408, New York, NY, USA. Association for Computing Machinery.
- Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. [How to do human evaluation: A brief introduction to user studies in NLP](#). *Natural Language Engineering*, 29(5):1199–1222.
- Anna Shadrova. 2021. [Topic models do not model topics: epistemological remarks and steps towards best practices](#). *Journal of Data Mining & Digital Humanities*, 2021.
- Samer Abdulateef Waheeb, Naseer Ahmed Khan, and Xuequn Shang. 2022. [Topic modeling and sentiment analysis of online education in the covid-19 era using social networks based datasets](#). *Electronics*, 11(5).

Hanna Megan Wallach. 2008. *Structured topic models for language*. Ph.D. thesis, University of Cambridge Cambridge, UK.

Elissa Nakajima Wickham and Emily Öhman. 2022. *Hate speech, Censorship, and Freedom of Speech: The Changing Policies of Reddit*. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Susik Yoon, Yu Meng, Dongha Lee, and Jiawei Han. 2023. *Scstory: Self-supervised and continual online story discovery*. In *Proceedings of the ACM Web Conference 2023*, pages 1853–1864, New York, NY, USA. Association for Computing Machinery.

A Detailed Participant Demography

The detailed participant demography is given in Table 4.

	N	Percentage
Continent		
Europe	45	75%
Africa	9	15%
Asia & Oceania	4	7%
North America	2	3%
Age		
18-29	41	68%
30-39	10	17%
40-49	6	10%
50+	3	5%
Education		
High School	15	25%
Technical college	8	13%
Undergraduate	21	35%
Graduate	14	23%
Doctorate	2	3%
Total		
Overall	66	100%
Rejected	6	9%
Used	60	91%

Table 4: Demography of the participants.

B Survey Details

B.1 Consent

Welcome!

Thank you for participating in this study! Your input is helping us develop transparent ways of evaluating AI.

AI models can be used to organize huge amounts of text documents. To the human eye it is not always obvious which features of the texts an AI model has cared about. Hence we see a need for a practical method for humans to evaluate how AI models organize texts.

For this study such AI models have been applied to group news articles. Your role in this study is to test an evaluation tool on these groups of news articles, and to assess whether the groups make sense. You are given 10 such groups for evaluation. The articles are in English and published in 2018 and 2022.

Your answers are anonymous and will not be used as training data for models. By agreeing to participate in this study, you consent to have your anonymous responses be stored at Umeå University and included in any research paper using this data.

For further questions or comments, contact the principal investigator Anton Eklund at anton.eklund@cs.umu.se.

B.2 Instructions to participants

B.2.1 UF

Introduction The evaluation tool is based on the assumption that the texts in each group have something in common. We call this a unifying feature. Unifying features are often Themes (e.g., Football, Politics, Natural Disasters, a person, a city), or Events (e.g., a particular football match, an election, a hurricane). Sometimes, the unifying feature is something else (e.g., opinions, formality) which are more difficult to identify but may still give insight into the AI models.

Your task is to explore the group and assess if any unifying feature exists. The models may have made mistakes when grouping the texts. If so, you should remove some texts from the group and define the feature from the remaining texts. You will be given the opportunity to explain why some texts are removed.

Explore the Group:

- Read all the titles carefully.
- Click on the title to show the text body of an article if you are unsure what it is about.
- Make up your mind what is most unifying for these articles.
- Exclude articles that don't fit with the rest of the group.

Unifying Feature:

- Select one of the following unifying feature types:

- Event: Something placed in time (e.g., a particular football match, an election, a hurricane).
 - Theme: Not bound by time (e.g., Football, Politics, Natural Disasters, a person, a city).
 - Other: Something unites the articles but is not an Event or a Theme (e.g., language style, opinion, formality).
- Name the group.
 - Optionally write a comment about why you are excluding some articles.

Rate the complexity of the task:

- You are asked to rate your agreement to statements about your experience assessing the group.

B.2.2 FT

Introduction The evaluation tool is based on the assumption that the texts in each group have something in common. It could be a theme (e.g., Football, Politics, Natural Disasters, a person, a city), or an event (e.g., a particular football match, an election, a hurricane). Sometimes, it is something else (e.g., opinions, formality) which may be more difficult to identify but will still give insight into the AI models.

Your task is to explore the group and assess whether the articles have anything in common. The models may have made mistakes when grouping the texts. If so, you should remove some texts from the group and name what is common in the remaining texts.

Explore the Group:

- Read all the titles carefully.
- Click on the title to show the text body of an article if you are unsure what it is about.
- Make up your mind what is most unifying for these articles.
- Exclude articles that don't fit with the rest of the group.

Name the group:

- Write a descriptive title of the group. Examples: "Football", "Eurovision Song contest 2022", "First-person stories"

Rate the complexity of the task:

- You are asked to rate your agreement to statements about your experience assessing the group.

B.2.3 TAX

Introduction The evaluation tool is based on the assumption that the texts in each group have something in common. It could be a theme (e.g., Football, Politics, Natural Disasters, a person, a city), or an event (e.g., a particular football match, an election, a hurricane). Sometimes, it is something else (e.g., opinions, formality) which may be more difficult to identify but will still give insight into the AI models.

Your task is to explore the group and assess whether the articles have anything in common. The models may have made mistakes when grouping the texts. If so, you should remove some texts from the group and name what is common in the remaining texts.

Explore the Group:

- Read all the titles carefully.
- Click on the title to show the text body of an article if you are unsure what it is about.
- Make up your mind what is most unifying for these articles.
- Exclude articles that don't fit with the rest of the group.

Define the group:

- Select one of the following general categories that you think the group fits under:
 - Culture
 - Entertainment
 - Politics
 - Crime
 - War
 - Lifestyle
 - Science
 - Home & Garden
 - Sports
 - Business & Finance
 - Personal Finance
 - Automotive
 - Weather

- Technology
 - Environment
 - Real Estate
 - Other
- Give a descriptive name to the group.

Rate the complexity of the task:

- You are asked to rate your agreement to statements about your experience assessing the group.

Empowering Teachers with Usability-Oriented LLM-Based Tools for Digital Pedagogy

Melany Macías, Lev Kharlashkin, Leo Einari Huovinen and Mika Hämäläinen

Metropolia University of Applied Sciences

Helsinki, Finland

first.last@metropolia.fi

Abstract

We present our work on two LLM-based tools that utilize artificial intelligence and creative technology to improve education. The first tool is a Moodle AI plugin, which helps teachers manage their course content more efficiently using AI-driven analysis, content generation, and an interactive chatbot. The second one is a curriculum planning tool that provides insight into the sustainability, work-life relevance, and workload of each course. Both of these tools have the common goal of integrating sustainable development goals (UN SDGs) into teaching, among other things. We will describe the usability-focused and user-centric approach we have embraced when developing these tools.

1 Introduction

In recent years, there has been an increase in the use of artificial intelligence (AI) in education that offers novel ways to support teachers and improve students' learning experiences (see Cleland Silva and Hämäläinen 2024). This shift has been particularly transformative in the field of digital humanities, where AI tools are being used to analyze vast datasets of historical texts (Hill and Hengchen, 2019; Partanen et al., 2021), digitize and interpret cultural artifacts (Benardou et al., 2018; Orlandi and Marsili, 2019) and even create interactive learning environments (Uibo et al., 2015; Ayeni et al., 2024).

AI-driven platforms now enable students to explore complex topics like literature, history, and philosophy through personalized learning paths (see Deng et al. 2024). Large language models (LLMs) in particular are now in the forefront of enabling the use of AI from a pedagogically insightful perspective (see Cleland Silva and Hämäläinen 2024). However, their full potential in education is yet to be capitalized as there are many obstacles such as a lack of understanding of the potential of LLMs (see Silmälä 2024) and legal issues such

as GDPR concerns that hinder universities from embracing them (see Hämäläinen 2024).

In this article, we present two LLM powered tools that show some of the ways LLMs can be used in educational settings. The first tool is a Moodle AI plugin, which integrates LLM capabilities within the Moodle¹ educational platform to assist educators in course creation and management. The second tool aims for curriculum planners to analyze the contents of course descriptions and degrees to better align them with the university's strategic goals.

Sustainability is an overarching theme for both of the tools as UN Sustainable Development Goals (SDGs)² are strategically important to the university. Recently, there has been a rise in interest in SDGs both in digital humanities (Smith et al., 2021; Fitri et al., 2021; Geraldo et al., 2023) and NLP (Amel-Zadeh et al., 2021; Matsui et al., 2022; Kharlashkin et al., 2024). The convergence of digital humanities and NLP in this context not only supports the advancement of sustainable development but also enriches the academic discourse around the role of technology in addressing global challenges.

This paper will provide a comprehensive overview of these tools, discussing their design and functionality. Furthermore, we have gathered user feedback through user studies and workshops. In the later sections of the paper, we describe the new features that we implemented based on the feedback. When building tools for educators, it is important to listen to user needs so that the tools will end up as useful as possible. This design philosophy is by no means new to the field of NLP (see Alnajjar et al. 2020).

¹<https://moodle.org/?lang=en>

²<https://sdgs.un.org/goals>

2 Related work

Ever since the emergence of ChatGPT (OpenAI, 2022), there has been an ever growing body of research on applying LLMs on every field of science. Digital humanities and digital pedagogy have not gone unaffected by this hype of the transformational new technology. In this section, we describe some of the LLM-based tools developed to empower teachers and curriculum planners.

This proliferation of education data, encompassing unstructured and structured text, relational databases etc. presents a complex challenge for constructing Knowledge Graphs (KGs). Recent research (Bui et al., 2024) has proposed methods for automatically building KGs from these heterogeneous data sources and explored the application of these KGs in conjunction with LLMs for improving the accuracy and relevance of responses in educational question-answering tasks.

Recent research (Hu et al., 2024) has begun to explore the role of LLMs in instructional design, particularly within the context of mathematics education. The study focused on assessing the instructional design capabilities of LLMs, using an instructional framework based on mathematical problem chains. The results indicated that LLMs, particularly GPT-4, performed well in various aspects of instructional design, such as setting instructional objectives, organizing teaching activities and selecting appropriate methods and strategies.

CodeAid (Kazemitabaar et al., 2024) is an LLM-based assistant designed to provide support without directly revealing code solutions. The deployment of CodeAid in a programming course with 700 students revealed key design considerations for future educational AI assistants, emphasizing the importance of leveraging AI's unique capabilities, promoting cognitive engagement through simplified query formulation, avoiding direct answers to encourage deeper learning and ensuring transparency to enable students to guide and assess AI responses.

Another contemporary research (Kahl et al., 2024) has explored the potential of LLMs as AI-based tutors in university courses. The study used standard similarity metrics such as BLEU-4, ROUGE, and BERTScore, alongside a human evaluation of helpfulness and trustworthiness, to assess the performance of the models. The findings suggest that combining RAG with prompt engineering significantly improves the factual accuracy of LLM-generated responses, making it an ideal ap-

proach for educational settings where additional course material can be leveraged to enrich model input.

3 Moodle AI Plugin

The purpose of the tool is to help teachers incorporate SDGs into their teaching. The functionalities include analysis of relevant SDGs, assignment suggestions and an interactive chatbot. The plugin can also create an executive summary of a course and highlight the most important skills and learning outcomes.

The idea behind developing a Moodle plugin instead of a separate system was motivated by the simple design principle of reducing the required effort from the user's perspective. When teachers can use the AI tool directly in a familiar platform, they do not need to learn an entirely new system, but instead interact with a new functionality in a more accessible way.

The plugin is implemented as injected JavaScript, PHP, and CSS within Moodle. This architecture allows for seamless integration with Moodle's existing infrastructure. When a user interacts with the plugin for the first time, a background process is triggered. Specifically, a post request is sent to a Flask server, where data scraping, preprocessing, and AI interactions take place.

During this initial interaction, a "super user" with managerial rights in Moodle scrapes the necessary data, such as slides. The data undergoes preprocessing stages that include language detection, removal of private information, and content refinement using the UralicNLP library (Hämäläinen, 2019). This preprocessing ensures that no personally identifiable information (PII) is passed to the AI model, ensuring full GDPR compliance. The removal of URLs, email addresses, and references, alongside the detection and elimination of named entities, guarantees that the data processed by the model is entirely anonymized, therefore the plugin strictly adheres to GDPR regulations and prioritizes data privacy throughout the process.

The processed information is then stored in the Chroma vector database³ as embeddings. These embeddings serve as the basis for various AI functionalities within the plugin. Lastly, the plugin uses the preprocessed slides and respective queries to prompt Google's PaLM 2 model (Anil et al., 2023)

³<https://www.trychroma.com/>

Summary

The slide series delves into the intricacies of embedded systems, their programming challenges, timing requirements, bit manipulation techniques, memory-mapped devices, IO interfaces, GPIO pin configurations, programmed IO techniques, and switch bounce filter mechanisms. It emphasizes the significance of reliability, fault tolerance, and safety in embedded systems, highlighting their distinct characteristics from general-purpose computers. The content underscores the importance of efficient resource allocation and constrained environments in embedded system design.

Related SDGs



Figure 1: Moodle AI Plugin main screen

over Vertex AI⁴. This model was chosen for its strong performance, cost and GDPR compliance, which aligns with the university's data privacy policies and its use of Google Cloud infrastructure. The plugin interacts with the model primarily through prompt-based requests, which enable the generation of summaries, related SDGs, suggestions of assignments or quizzes based on course content or conversations with the chatbot.

To optimize performance, the plugin employs a caching mechanism using a MySQL database. Once the slides for a course have been processed, the results are stored in the database, allowing the plugin to serve future requests without contacting the Flask server again.

The Moodle AI plugin currently supports English and Finnish. We implemented localization, allowing the plugin to automatically switch languages based on Moodle's language settings. For instance, the string `$string['pluginname']` is defined as 'AI Summarizer' in English and 'AI Yhteenveto' in Finnish. This logic ensures that every text element in the plugin is fully translatable. Future language expansion will be straightforward, requiring only the addition of translations, as the localization system is already in place.

The plugin's first screen includes a summary of slides and the SDGs related to that content. When the user clicks on one SDG, a new window opens

with more information about that SDG containing an explanation of how it relates to the slides, assignment suggestions and a draft speech that would help teachers mention the sustainability aspect during their lectures.

Another functionality included in the plugin is the chatbot, which makes use of the embeddings stored on the Chroma database along with Vertex AI to implement a retrieval-augmented generation (RAG) system (Lewis et al., 2020). The RAG system pulls relevant content from the Chroma vector database, which stores embeddings of course slides and related metadata, and combines this with AI-generated responses. LangChain⁵ is used to manage conversational history and memory, ensuring a structured conversation flow that enables the chatbot to provide context-aware, accurate answers across multiple interactions. This integration allows the chatbot to assist teachers by offering personalized, relevant responses that are directly grounded in the specific course material.

Although finetuning has not yet been implemented in this iteration of the chatbot, future improvements, as outlined in (Macías, 2024), might involve fine-tuning the model to enhance its performance for specific educational tasks.

Additionally, the plugin includes the functionality of analyzing all the slides of the same course. This system relies on metadata stored in the

⁴<https://cloud.google.com/vertex-ai?hl=en>

⁵<https://python.langchain.com/>

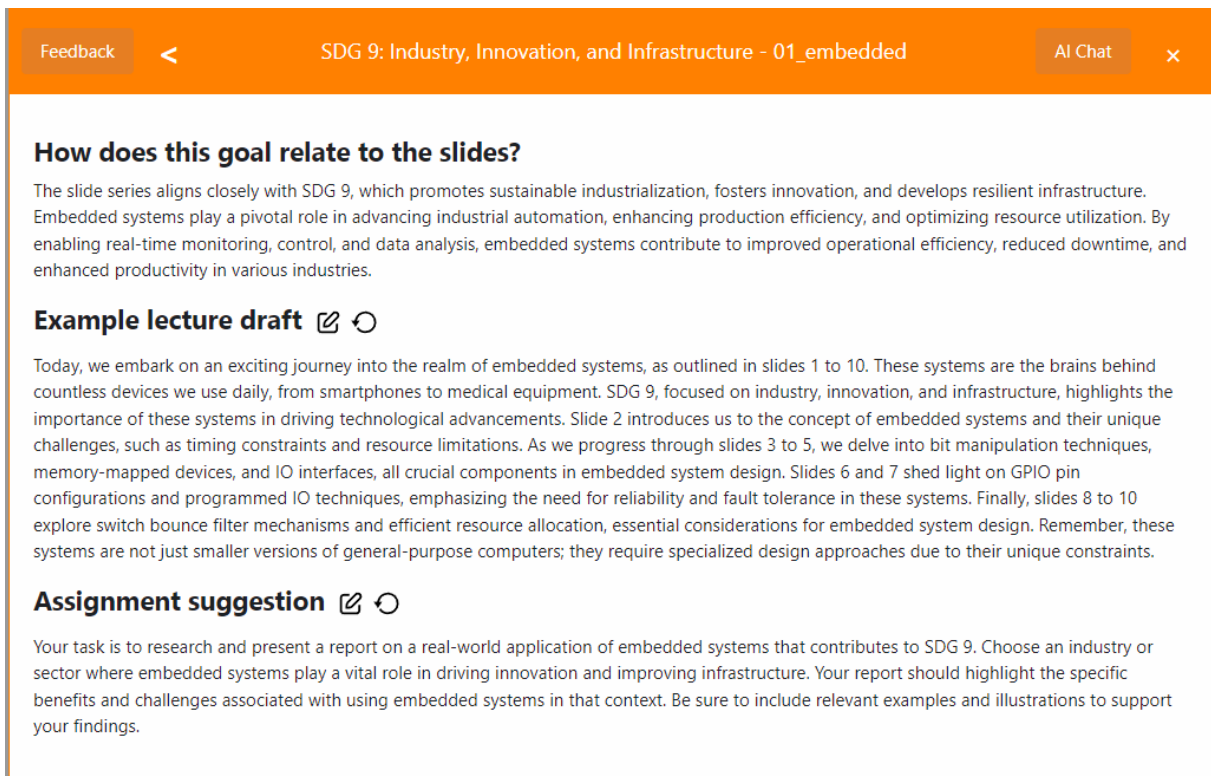


Figure 2: Moodle AI Plugin specific SDG’s information

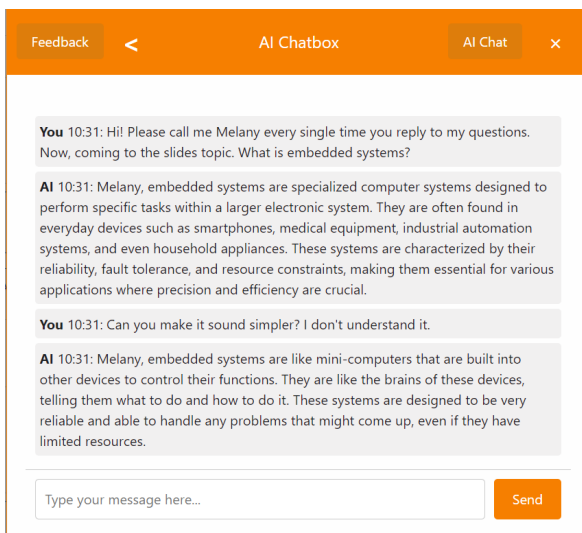


Figure 3: Moodle AI Plugin chatbot

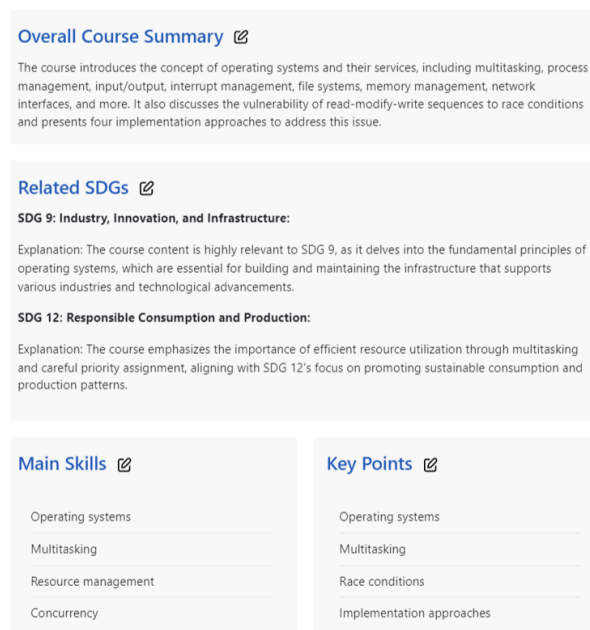


Figure 4: Moodle AI Plugin overall course functionality

Chroma vector database, where each slide is linked to a specific course through its course ID. The retrieved content is then fed into the AI model to generate a summary, key points, main skills, and related SDGs.

The overall course functionality also includes a general chatbot that is designed to answer queries related to the course. The chatbot, like other parts of the plugin, utilizes a RAG system, but it focuses

exclusively on the content from the whole course. This ensures that the chatbot’s responses are always relevant to the specific course material, providing a personalized experience.

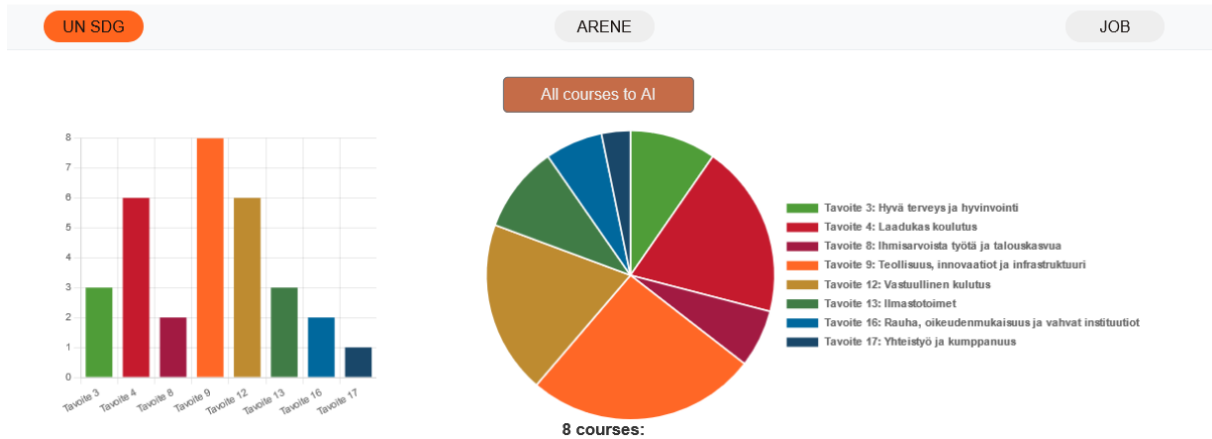


Figure 5: The curriculum analyzer calculating the SDGs of a degree

4 Curriculum planning tool

The curriculum planning tool is a separate tool that operates on a different level of abstraction to the Moodle plugin. This tool works directly on course descriptions and degree structures. The functionalities include SDG analysis, analysis of Arene competences⁶ and work life relevance analysis.

On the technical side, this tool uses PaLM 2 as well. The tool's architecture separates the AI processing, which is handled server-side through Flask and Vertex AI, from the sensitive curriculum data and interface, managed in MariaDB and React. This division between local and server-side components also allows the tool to offload computationally intensive tasks to the server while maintaining a responsive user interface. The server-side components, including the PaLM 2 model accessed through Vertex AI, handle the core analysis tasks, ensuring that sensitive curriculum and course data remain secure. This allows for powerful AI analysis while keeping the human user in control of the interpretation and application of results.

The course-level SDG analysis feature directly addresses a critical gap in current curriculum planning tools: the lack of systematic methods for integrating sustainability principles across educational programs. While many institutions recognize the importance of sustainability education, there has been a dearth of practical tools to support this integration at the course level (Lozano et al., 2013; Aurandt and Butler, 2011). This tool provides educators with actionable insights into how individual

courses contribute to sustainability goals, facilitating more informed curriculum design decisions.

The methodology employed in degree-level SDG analysis typically involves aggregating and synthesizing data from individual course assessments to create a program-wide overview. Integration of SDGs into this mapping process represents a novel extension of traditional curriculum mapping techniques.

The visual representation of SDG alignment across all the courses of any one degree, using charts and color-coded matrices, offers an intuitive interface for understanding complex relationships between course content and sustainability objectives. This visual approach addresses the need for more accessible and interpretable data in curriculum planning, a gap that has been noted in educational technology research.

The incorporation of work-life competencies analysis in curriculum planning tools addresses the growing demand for alignment between higher education outcomes and labor market needs. This feature responds to the call for more employability-focused curriculum design, as emphasized by Pereira et al. (2020) in their stakeholder analysis of university curricula and employability.

5 Workshop and Feedback

We organized three workshops, each lasting 2 hours, to gather feedback from educators on the usability and features of the Moodle plugin. The number of participants varied between 5 and 10 for each workshop, with attendees representing different departments within the university. Two of the workshops were held in person in different cam-

⁶<https://arene.fi/julkaisut/suositus-ammattikorkeakoulujen-yhteisista-kompetensseista-janniiden-soveltamisesta-2022/>



Figure 6: Updated Moodle AI Plugin

uses, whereas one workshop was organized online over Teams. Furthermore, we interviewed study planners on their needs for the curriculum tool (see [Huovinen 2024](#)).

As a result of the workshops, several usability problems were identified such as teachers losing track of where they were in the plugin, some functionalities not working on Firefox, our regular expressions failed parsing the output of the LLM and so on. Additionally, teachers were not happy with the quality of the generated assignments as they wanted more control over how broad the assignments were as well as being able to generate quizzes automatically.

Most importantly, the teachers emphasized the need for a tool whose focus is assisting with the teaching experience in general and not specifically just with sustainability.

In response to this feedback, we shifted our approach from having sustainability as the primary focus to creating a more versatile tool that supports teaching in a broader context. While sustainability remains an integral part of the plugin, we have expanded its functionalities to better assist teachers in various aspects of their courses. The updated design, featuring a hamburger menu and separate windows for different functionalities, reflects this shift.

Now, the plugin offers tools for generating quizzes, assignments, and interactive content, alongside its original sustainability features and chatbot capabilities. In the updated plugin, each functionality is separated, ensuring that teachers don't lose track of where they are in the process. The new layout not only enhances usability but also ensures that the plugin's functionalities are more accessible.

The new Quizzes functionality allows teachers to automatically generate quizzes based on the content of their slides. Teachers can select the number of questions and the question type, such as multiple-choice or true/false. Additionally, teachers can input specific guidelines or constraints for the quiz generation, ensuring that the generated quizzes align with their teaching goals.

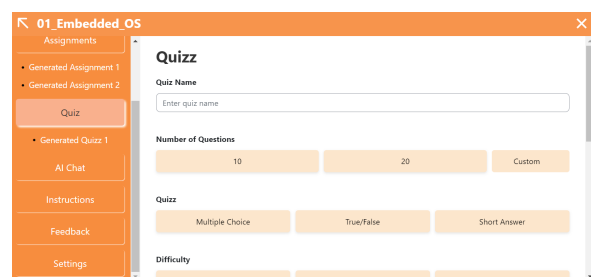


Figure 7: Updated Moodle AI Plugin quizzes

The Quizzes window interacts with the Moodle backend to create actual quizzes within the Moodle platform, making it easy for teachers to integrate the quizzes into their courses. The AI takes into consideration the teacher’s suggestions, ensuring that the generated content meets their expectations.

Additionally, the new Assignment Generator functionality allows teachers to create assignments tailored to their course content. Teachers can select various parameters such as the type of assessment (e.g., projects, essays, presentations), the difficulty level, and the content focus. Similar to the Quizzes functionality, teachers can provide specific instructions or constraints to guide the assignment generation process.

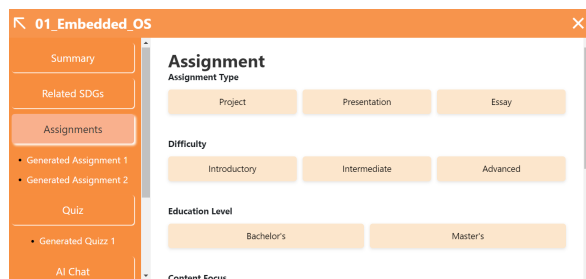


Figure 8: Updated Moodle AI Plugin assignments

The AI uses the input provided by teachers to generate assignments that align with the course content and the teacher’s objectives. This feature was designed to give teachers more control over the generated content, addressing the feedback we received during the workshops about the need for more customizable assignments.

As for the curriculum planner, there was a demand for a workload assessment functionality and quality assessment functionality. These functionalities require more future work on integrating the Moodle AI plugin with the curriculum tool. As this requires a more systematic approach to analysing the material of all of our teachers, this deeper integration has reached a state of hiatus due to organizational resistance.

6 Conclusions

In this paper, we have presented our hands-on tools that bring LLMs closer to educators to empower a digital transformation of pedagogy at our university. By integrating these tools into the existing educational framework, we aim to enhance the teaching and learning experience, making it more interactive, personalized and efficient for all parties

involved. Furthermore, we have organized workshops to help instructors effectively utilize these technologies and to gather their feedback.

In addition to transforming pedagogy, our LLM-based tools pave the road for digital humanities research. By leveraging the capabilities of LLMs, educators and researchers can explore novel methods for analyzing and interpreting vast amounts of learning data. This can be used to further understand the interplay of LLMs, teachers and students in educational contexts.

7 Limitations

This study was conducted exclusively within one university of applied sciences (UAS) in Finland, which limits the generalizability of the findings. The specific context, policies and educational frameworks of Finnish higher education, particularly within this UAS, may not fully reflect the conditions in other countries or universities. As such, the results and conclusions may not scale to different educational systems or institutions with varying academic cultures, regulatory environments or student demographics.

References

- Khalid Alnajjar, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020. Ve’rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 1–6.
- Amir Amel-Zadeh, Mike Chen, George Mussalli, and Michael Weinberg. 2021. Nlp for sdgs: measuring corporate alignment with the sustainable development goals. *Columbia Business School Research Paper*.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Jennifer L Aurdant and Elizabeth C Butler. 2011. Sustainability education: Approaches for incorporating sustainability into the undergraduate curriculum. *Journal of Professional Issues in Engineering Education and Practice*, 137(2):102–106.
- Oyebola Olusola Ayeni, Nancy Mohd Al Hamad, Onyebuchi Nneamaka Chisom, Blessing Osawaru, and Ololade Elizabeth Adewusi. 2024. Ai in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, 18(2):261–271.

- Agiatis Benardou, Erik Champion, Costis Dallas, and Lorna M Hughes. 2018. *Cultural heritage infrastructures in digital humanities*. Routledge London.
- Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. 2024. Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut. In *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia*, pages 36–43.
- Tricia Cleland Silva and Mika Hämäläinen. 2024. Innovating for the future: Ai and hrm capabilities for sustainability in higher education. In *Academy of Management Annual Meeting*, volume 2024.
- Weifeng Deng, Lin Wang, and Xue Deng. 2024. Strategies for optimizing personalized learning pathways with artificial intelligence assistance. *International Journal of Advanced Computer Science & Applications*, 15(6).
- Hidayatul Fitri, Widyawan Widyawan, and Indah Soesanti. 2021. Topic modeling in the news document on sustainable development goals. *IJITEE (International Journal of Information Technology and Electrical Engineering)*, 5(3):82–89.
- Genilson Geraldo, Edgar Bisset-Alvarez, and Marli Dias de Souza Pinto. 2023. Digital humanities and the sustainable development goals: a reflection for information science. *Transformação*, 35:e227210.
- Mika Hämäläinen. 2024. Legal and ethical considerations that hinder the use of llms in a finnish institution of higher education. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies@ LREC-COLING 2024*, pages 24–27.
- Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.
- Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design. *IEEE Transactions on Learning Technologies*.
- Leo Huovinen. 2024. Assessing usability of large language models in education. *Metropolia University of Applied Sciences, BSc thesis*.
- Mika Hämäläinen. 2019. [UralicNLP: An NLP library for Uralic languages](#). *Journal of Open Source Software*, 4(37):1345.
- Sebastian Kahl, Felix Löffler, Martin Maciol, Fabian Ridder, Marius Schmitz, Jennifer Spanagel, Jens Wienkamp, Christopher Burgahn, and Malte Schilling. 2024. Evaluating the impact of advanced llm techniques on ai-lecture tutors for a robotics course. *arXiv preprint arXiv:2408.04645*.
- Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.
- Lev Kharlashkin, Melany Macias, Leo Huovinen, and Mika Hämäläinen. 2024. Predicting sustainable development goals using course descriptions—from llms to conventional foundation models. *Journal of Data Mining and Digital Humanities*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Rodrigo Lozano, Francisco J. Lozano, Karel Mulder, Donald Huisingh, and Tom Waas. 2013. [Advancing higher education for sustainable development: international insights and critical reflections](#). In *Journal of Cleaner Production*, volume 48, pages 3–9, Mexico City, Mexico.
- Melany Macías. 2024. [Finetuning and improving prediction results of llms using synthetic data](#). Bachelor’s Thesis, Metropolia University of Applied Sciences.
- Takanori Matsui, Kanoko Suzuki, Kyota Ando, Yuya Kitai, Chihiro Haga, Naoki Masuhara, and Shun Kawakubo. 2022. A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders. *Sustainability Science*, 17(3):969–985.
- OpenAI. 2022. [Introducing chatgpt](#). *OpenAI Blog*.
- Lucia Maria Orlandi and Giulia Marsili. 2019. Digital humanities and cultural heritage preservation. *Studies in Digital Heritage*, 3(2):144–155.
- Niko Partanen, Jack Rueter, Khalid Alnajjar, and Mika Hämäläinen. 2021. Processing ma castrén’s materials: Multilingual historical typed and handwritten manuscripts. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 47–54.
- Elisabeth T Pereira, Madalena Vilas-Boas, and Cátia FC Rebelo. 2020. University curricula and employability: The stakeholders’ views for a future agenda. *Industry and higher education*, 34(5):321–329.
- Petri Silmälä. 2024. Tekoälyn sietämätön keveys. *Hiiltä ja Timanttia*.

Thomas Bryan Smith, Raffaele Vacca, Luca Mantegazza, and Ilaria Capua. 2021. Natural language processing and network analysis provide novel insights on policy and scientific discourse around sustainable development goals. *Scientific reports*, 11(1):22427.

Heli Uibo, Jaak Pruulmann-Vengerfeldt, Jack Rueter, and Sulev Iva. 2015. Oahpa! opi! opiq! developing free online programs for learning estonian and voro. In *Proceedings of the fourth workshop on NLP for computer-assisted language learning*, pages 51–64.

Author Index

- Adam, Raven, 75
Alex, Beatrice, 319
Alnajjar, Khalid, 499
Amano, Kyoko, 123
Anderson, Carolyn Jane, 17
Aygen, Gülşat, 352
- Bagdasarov, Sergei, 483
Bando, Ryo, 331
Barbosa, Denilson, 85
Bārzdīņš, Guntis, 289
Baunvig, Katrine, 140
Biemann, Chris, 423
Björck, Henrik, 280
Brown, Susan, 85
Byun, Courtni, 63
- Chatterjee, Anindita, 156
Chopra, Medini, 156
Clerice, Thibault, 189
Cruciani, Giulia, 41
- Dargis, Roberts, 289
Das, Partha Pratim, 156
Deichsler, Florian, 247
Dent, Rasul, 189
Dey, Lipika, 156
Dipper, Stefanie, 177
Drewes, Frank, 536
Du, Quanqi, 113
- Edalat, Abbas, 259
Eklund, Anton, 536
Enevoldsen, Kenneth, 140
Erten-Johansson, Selcen, 308
Eskelinen, Anni, 308
- Feldkamp, Pascale, 140
Feng, Yu, 200
Fischer, Tim, 423
Forsman, Mona, 536
Freedman, Reva, 352
Freyermuth, Elian, 386
- Gorovaia, Svetlana, 398
- Hämäläinen, Mika, 499, 549
HANNANI, Mohamed, 489
Hashimoto, Yuta, 331
Hellström, Saara, 308
Henderson, James, 513
Henriksson, Erik, 216, 308, 386
Hervieux, Natalie, 85
Higashiyama, Shohei, 331
Hoste, Veronique, 113
Hou, Xinmeng, 362
Huovinen, Leo Einari, 549
- Igarashi, Ryo, 413
Inoshita, Keito, 504
- janes, juliette, 189
Jang, Woori, 339
Jenkins, Chris W., 240
Jung, Seohyon, 339
- Kanerva, Jenna, 216
Kanner, Antti, 386
Kardos, Márton, 140
Katayama, Ayuki, 331
Kenneth, Mary Ogbuka, 259
Kharlashkin, Lev, 549
Khosmood, Foaad, 259
Khurshudyan, Victoria, 438
Kostkan, Jan, 140
Kousa, Ilona, 521
Krusic, Lucija, 51
Kupari, Hanna-Mari Kristiina, 216
Kurzynski, Maciej, 200
Kyogoku, Yuki, 123
- Laarmann-Quante, Ronja, 177
Laippala, Veronika, 216, 308, 386
Lamb, William, 319
Lassche, Alie, 140
Li, Lucian, 9
Liimatta, Aatu, 1
Lin, Chenghua, 450
Lingras, Pawan, 377
Lippincott, Thomas, 325

Macias, Melany Vanessa, 549
Mago, Vijay, 377
Manrique-Gomez, Laura, 132
Manrique, Ruben, 132
Marfurt, Andreas, 513
Marivate, Vukosi, 229
Maynard, Diana, 450
Meaney, J. A., 319
Mervaala, Erkki, 521
Messner, Craig, 325
Miletić, Filip, 240
Miyagawa, So, 123, 413, 463
Montes, Tony, 132
Myntti, Amanda, 308, 386

Neveditsin, Nikita, 377
Nguyen, Thong Thanh, 165
Nguyen, Truc-My, 165
Nielbo, Kristoffer, 140

Ogiso, Toshinobu, 331
Ohlsson, Claes, 280
Ohman, Emily Sofi, 1
Ohmukai, Ikki, 275
Ortiz Suarez, Pedro, 189
Ouchi, Hiroki, 331

Pawłowski, Adam, 105
Peng, Bohua, 450
Piper, Andrew, 472

Rajaei Moghadam, Mina, 352
Rassem, Malak, 240
Repo, Liina, 308, 386
Rezaei, Mosab, 352
Roberts, Jesse, 31
Roberts, Lindsey, 31
Robinson, Ambrose, 450
Rodriguez Herrera, Arturo, 132
Roussel, Adam, 209
Rubinstein, Aynat, 294
Rueter, Jack, 499
Runefelt, Leif, 280
Ruths, Derek, 472

Sagot, Benoît, 189
Sakai, Yusuke, 331
Salgaonkar, Ambuja, 377
Saulite, Baiba, 289
Schmidt, Gleb, 398
Schulte im Walde, Sabine, 240
Schulz, Konstantin, 247
Seppi, Kevin, 63

Shmidman, Avi, 294
Sindane, Thapelo Andrew, 229
Skadiņa, Inguna, 289
Soudi, Abdelhadi, 489
Sylvan, David, 513

Tahmasebi, Nina, 280
Takeuchi, Ayano, 331
Teich, Elke, 483
Thorne, William, 450
Thornton, Ashley, 513
Tomeh, Nadi, 438
Tsigkouli, Myrto, 240
Tsukagoshi, Yuzuki, 123, 275

Umphrey, Ray, 31

Van Laerhoven, Kristof, 489
Vasicek, Piper, 63
Venglarova, Klara, 75
Vidal-Gorène, Chahan, 438
Virk, Shafqat Mumtaz, 280
Vogeler, Georg, 75

Walkowiak, Tomasz, 105
Wannaz, Audric-Charles, 463
Watanabe, Taro, 331

Xu, Michael, 472
Xu, Xiaotong, 200

Yamshchikov, Ivan P., 398
Yang, Funing, 17
Yao, Peiran, 85